European Research Council

ERC Advanced Grant 2010
research proposal (Part B1)

Modeling, interpreting and manipulating digital video

VIDEOWORLD

- Name of PI: Jean Ponce
- Host institution: INRIA
- Proposal full title: Modeling, interpreting and manipulating digital video
- Proposal short name: VideoWorld
- Proposal duration in months: 60 months

Proposal summary: Digital video is everywhere, at home, at work, and on the Internet. Yet, effective technology for organizing, retrieving, improving, and editing its content is nowhere to be found. Models for video content, interpretation and manipulation inherited from still imagery are obsolete, and new ones must be invented. With a new convergence between computer vision, machine learning, and signal processing, the time is right for such an endeavor. Concretely, we will develop novel spatio-temporal models of video content learned from training data and capturing both the local appearance and nonrigid motion of the elements—persons and their surroundings—that make up a dynamic scene. We will also develop formal models of the video interpretation process that leave behind the architectures inherited from the world of still images to capture the complex interactions between these elements, yet can be learned effectively despite the sparse annotations typical of video understanding scenarios. Finally, we will propose a unified model for video restoration and editing that builds on recent advances in sparse coding and dictionary learning, and will allow for unprecedented control of the video stream. This project addresses fundamental research issues, but its results are expected to serve as a basis for groundbreaking technological advances for applications as varied as film post-production, video archival, and smart camera phones.

1. The principal investigator

1a. Scientific leadership profile

I received the "doctorat de troisième cycle" and "doctorat d'état" degrees in computer science from the University of Paris XI in 1983 and 1988. After over 20 years in the US, first as a research scientist at MIT and Stanford, then as a faculty member at the University of Illinois at Urbana-Champaign (UIUC), I returned to Europe in 2006 to join the faculty of Ecole Normale Supérieure (ENS) in Paris. In 2007, I started a joint ENS/INRIA research team, called Willow, dedicated to computer vision and machine learning. Willow has quickly reached maturity, with about 30 members and a well-established international reputation. It is about to spin off its core machine learning activities, wich will give me an opportunity, with a reduced administrative load, to focus the next five years of my own research on the visual analysis of video data, a project with many scientific challenges but the potential for groundbreaking research and great societal impact.

Content and impact of major scientific contributions: My main research interest is in computer vision, the area of computer science and engineering devoted to the automated computer interpretation of digital imagery ("what is the depth of this pixel in the scene?", "is there a chair in this image?"). Much of my early work was dedicated to shape representation, with the first answers to a number of open problems, including the first effective algorithm for computing differential invariants of surfaces from range data [CVGIP'85, 264 citations],[1] the first formal proof of the existence of viewpoint invariants for generalized cylinders [PAMI'89, 128 citations]; the first implemented algorithms for computing the 3D pose of a solid bounded by a curved surface from a photograph [PAMI'90, 202 citations] and for computing the aspect graph of such a solid [IJCV'90, 110 citations] (this had been an open problem since 1976), and the first computational characterization of compact solids bounded by algebraic surfaces [PAMI'94, 134 citations].

In the mid-90s, I made a foray into sensorless, geometric robotics (grasp and manipulation planning). This includes two papers on the construction of stable, "form-closure" grasps for polygons [TRA'95, 179 citations] and polyhedra [IJRR'97, 175 citations] that are now classical references in the field. This work also resulted in the design and construction of several grippers and manipulation devices, as well as a US patent (2003). At the same time, I wrote "Computer Vision: A Modern Approach" (with D. Forsyth, Prentice-Hall, 2002, 1801 citations), which is now the leading computer vision textbook worldwide: It is used at CMU, Oxford, MIT, and UC Berkeley for example, has sold over 10,000 copies, and has been translated in Chinese, Japanese, and Russian.

After completing this book, I returned full-time to computer vision research, focusing on three fundamental problems before leaving UIUC: 3D object recognition, including an algorithm [3, 101 citations] whose implementation is publicly available under an open-source license and has been transfered to Bertin Technologies and Toyota; category-level image classification and object detection, including the widely influential *spatial pyramid* approach [1, 540 citations]; and 3D photography, including the recent (2007) *PMVS* algorithm [6, 60 citations] for multi-view stereo reconstruction that is generally acknowledged as the most accurate to date (winner in 4 of the 6 categories of the Middlebury competition `http://vision.middlebury.edu/mview/`), and is also publicly available under an open-source license.

[1] In this section, papers are refered to by the name of the journal/conference where they appear, followed by the year of their publication. Numbered references correspond to the corresponding publications in the "top 10" list of Section 1.c. All citation numbers are from *Google Scholar*. Acronyms: AR = Advanced Robotics, CVGIP = Computer Vision, Graphics and Image Processing, CVPR = IEEE Conf. on Computer Vision and Pattern Recognition, ECCV = European Conf. on Computer Vision, FTCGV = Foundations and Trends in Computer Graphics and Vision, ICCV = Int. Conf. on Computer Vision, ICML = Int. Conf. on Machine Learning, IJRR = Int. Journal on Robotics Research, IROS = IEEE/RSJ Conf. on Intelligent Robots and Systems, ISRR = Int. Symp. on Robotics Research, JMLR = Journal of Machine Learning Research, NIPS = Neural Information Processing Systems, PAMI = IEEE Trans. on Pattern Analysis and Machine Intelligence, SJIS = SIAM Journal on Imaging Sciences, TRA = IEEE Trans. on Robotics and Automation.

By the time I returned to France in 2006, I realized two things: (1) with video everywhere, from family footage to TV or the Internet (ours is a video world!), the automated analysis of digital video is the future of computer vision; and (2) true interdisciplinary collaboration with machine learning researchers (as opposed to rote use of textbook classification techniques) is a prerequisite to further progress. This vision is implemented in the Willow research team, that successfully (and about equally) divides its activities between computer vision, machine learning, and the cross-pollination of the two fields, with video as one of the core research areas. With my recent work [ICCV'09, ICML'09, JMLR'10] on sparse coding and dictionary learning for image restoration, a third key idea has imposed itself: (3) a new alliance between computer vision, machine learning, and signal processing is afoot, and VideoWorld is a unique opportunity to jump start it and make Europe its leader.

International recognition and diffusion: I am the author of over 150 technical publications, and my h-number is 37, with over 6000 citations.[2] Four of my conference papers [IROS'97, ISRR'99, ECCV'06, CVPR'09] have been selected for publication in special issues of AR (1998), IJRR (2001), IJCV [9], and PAMI (2010) dedicated to the best (typically four to six) papers from these conferences. O. Duchenne, the first author on the CVPR'09 paper and one of my current PhD students, has received the "best student paper - honorable mention" award at that conference. My scientific leadership activities include chairing several international conferences (CVPR'97, CVPR'00, ECCV'08), organizing numerous workshops, and serving on the editorial board of several journals (CVIU, FTCGV, IJCV, SJIS, TRA). In particular, I served from 2003 to 2007 as editor-in-chief of IJCV, one of the top two journals of the field. Under my tenure, the impact factor of IJCV steadily grew from 2 to 6, and IJCV was ranked first of all journals in computer science (source: ISI Web of Knowledge) when I decided to step down. I am also an IEEE Fellow, have served on several scientific advisory boards for academia, government, and industry, and have received a US patent for the development of a robotic parts feeder.

Effort and ability to inspire younger researchers: I have supervised 10 PhD students at UIUC. All of them are enjoying successful careers in academia or industry. Notable among these are Ilan Shimshoni (PhD, 1995), who is a professor (and former department head) in the department of information management systems at Haifa University in Israel; Steve Sullivan (PhD, 1997), who is the director of research and development at Industrial, Light and Magic, the world leading visual effects company, and is the recipient of two Academy technical achievement awards; Attawith Sudsang (PhD, 1999), who decided to return to his native Thailand, and is now a faculty member in the department of computer engineering at Chulalongkorn University, the top academic institution there; Svetlana Lazebnik (PhD, 2006), who is now an assistant professor in the computer science department of the University of North Carolina at Chapel Hill, and a recipient of an NSF career award as well as a Microsoft research faculty fellowship; and Yasutaka Furukawa (PhD, 2007), who recently joined Google after a post-doctoral stay at the University of Washington. Lazebnik and Furukawa are the first authors of the *spatial pyramid* and *PMVS* papers I mentioned earlier, and they will be stars of the new computer vision generation. I am currently advising or co-advising 6 PhD students and 3 post-docs within Willow. The first one of my Willow PhD students, J. Mairal, will graduate this summer.

Proven ability to productively change research fields and/or establish new interdisciplinary approaches: I have demonstrated twice my ability to switch fields, first in the 1990s with my work in robotics (over 570 citations, one US patent), then in the past three years with my involvement in machine learning research through the creation of Willow and (so far) two papers in NIPS (2008) and ICML (2009), the top conferences of this field, and one article in JMLR (2010), its top international journal. My recent work on image denoising and demosaicking with J. Mairal and F. Bach [ICCV'09] is another foray into a new field, that of signal processing.

[2]The h-number measures scientific impact as the maximum number h of publications from a given author that have been cited at least h times. The h-number reported here has been computed using the popular *Harzing's Publish or Perish* software. Although I have made all efforts to manually eliminate duplicates and false citations, it is nearly impossible to account for all self citations, of which a small number likely remains.

1b. Curriculum Vitae

Education

• Ecole Normale Supérieure de l'Enseignement Technique, Cachan, Mathematics, 1978–1982.
• Doctorat de Troisième Cycle, Computer Science, University of Paris Orsay, 1983.
• Doctorat d'Etat, Computer Science, University of Paris Orsay, 1988.

Employment

• INRIA: Research scientist, 1982–1985.
• MIT Artificial Intelligence Laboratory: Visiting scientist, 1984–1985.
• Stanford University, Dept. of Computer Science: Research associate, 1985–1989. Sr. research associate, 1988–1989.
• University of Illinois at Urbana-Champaign (UIUC), Dept. of Computer Science and Beckman Institute: Asst. professor, 1990–1993. Assoc. professor (tenured), 1993–1998. Professor, 1998–2006.
• Ecole Normale Supérieure (ENS), Dept. of Computer Science: Professor, first class, 2006. Professor, exceptional class, 2007–.

Academic honors and awards

• Xerox award for faculty research, College of Engineering (CoE), UIUC (1993, 1998): This award is given annually to three assistant professors (Jr. award) and three associate professors (Sr. award) in the CoE, which counts 12 departments and is ranked in the top 5 in the US. I received both awards, in 1993 and 1998 respectively.
• Center of Advanced Study associate, UIUC (1994): This award is given annually to a dozen tenured faculty members campuswide. In 1994, I was one of its two recipients named "Beckman associate", which further recognizes outstanding achievements by young researchers.
• Outstanding undergraduate advisor award, CoE, UIUC, 2000.
• IEEE fellow, 2003.

Professional activities

• Area Editor, Computer Vision and Image Understanding, 1994–2000.
• Member, ARPA/ORD RADIUS Image Understanding advisory committee, 1994–1996.
• Member, ARO Computational Geometry for Intelligent Systems advisory board, 1996–1998.
• Associate editor, IEEE Transactions on Robotics and Automation, 1998–2001.
• Member, scientific advisory board of Electricité de France, 1998–2002.
• Member, scientific advisory board of France Télécom, 2001–2004.
• Editorial board member, International Journal of Computer Vision, 2001–.
• Member, network of North American advisors to the French Academy of Engineering, 2002–2006.
• Editor-in-chief, International Journal of Computer Vision, 2003–2007.
• Editorial board member, Foundation and Trends in Computer Graphics and Vision, 2005–.
• Member, scientific advisory board of the Institute of Ecole Normale Supérieure, 2007–.
• Member, contents thematic commission, Cap Digital, 2009–.
• Editorial board member, SIAM Journal on Imaging Sciences, 2009–.

Publications

Over 150 publications, including: 1 textbook, translated in three languages (2003); 3 edited collections; 15 book chapters; 45 journal articles; and 88 refereed conference papers.

US patent

• *Automated reconfigurable object manipulation device with an array of pins:* US Patent # 6,633,797. See [Akella, Blind, McCullough, Ponce, IJRR 20(10):808-818, 2001] for details.

Software

Several significant sotware packages developed in my research groups at UIUC and INRIA are available under open-source licenses:
• *3D recognition software:* This is a C implementation of the recognition method described in [3]. It is

avaible at: `http://www-cvr.ai.uiuc.edu/ponce_grp/software/3d.html`, and has been transfered to Bertin Technologies and Toyota.
• *PMVS:* This is a C implementation of the multi-view stereo algorithm described in [6]. It is available at: `http://grail.cs.washington.edu/software/pmvs/`.
• *SPAMS:* This is an optimization toolbox for efficient sparse coding and dictionary learning. See [Mairal, Bach, Ponce, Sapiro, Proc. ICML, 2009] for details. It is available at: `http://www.di.ens.fr/willow/SPAMS/`.

Some recent invited lectures (2006–2010)
2006: Mathematics and Image Analysis Conference, Paris, France; Johns Hopkins University; Rensselaer Polytechnic Institute CS Day; University of North Carolina, Chapel Hill. **2007:** ACCV'07 Vision Workshop, Hiroshima; ICCV'07 3D Vision Workshop, Rio de Janeiro; Microsoft Research, Cambridge. **2008:** European Workshop on Computational Geometry, Paris; Ecole Normale Supérieure de Cachan; Ecole Polytechnique, Paris; Microsoft Tech Days, Paris; Télécom Paris; International Workshop on Computer Vision, Venice; International Workshop on Shape Perception in Human and Computer Vision, ECCV'08; New York University, NYC. **2009:** Beckman Institute 20th Anniversary Symposium, Urbana; Laboratoire d'Analyse et d'Architecture des Systèmes, Toulouse; University of Southern California, Los Angeles. **2010:** Keynote speaker, British Machine Vision Conference, Aberystwyth, Wales; Janelia Conference on What Can Computer Vision Do for Neuroscience and Vice Versa, VA; Laboratoire d'informatique Gaspard-Monge, Paris; Laboratoire Jacques-Louis Lions, Paris.

Student and post-doc supervision
• **UIUC:** 8 MS students (graduated), and 10 PhD students (graduated): I. Shimshoni (1995, professor, MIS department, Univ. of Haifa), T. Joshi (1995, MSR Bangalore), S. Sullivan (1997, head of R&D at ILM), A. Sudsang (1999, assistant prof. at Chulalongkorn University, Thailand), Y. Genc (1999, Siemens SCR), F. Rothganger (2004, Sandia Labs), S. Lazebnik (2006, assistant professor, UNC Chapel Hill), K. McHenry (2008, National Center for Supercomputing Applications), A. Kushal (2008, Two Sigma Investments), Y. Furukawa (2008, Google).
• **Willow:** 6 PhD students (current): Y. Boureau, F. Couzinie-Devy, O. Duchenne, A. Joulin, J. Mairal, O. Whyte, and 3 post-docs (current): Kong H., B. Russell, J. van Gemert.

Teaching
• **UIUC:** Numerical methods (sophomores); introduction to computer graphics (seniors); advanced computer graphics (seniors); introduction to artificial intelligence (seniors); geometric and symbolic computation (seniors); geometric modeling (seniors); computer vision (graduate students); advanced robotic planning (graduate students); geometric methods in computer vision (graduate students).
• **ENS:** Introduction to scientific computing and its applications; geometry and computer vision; object recognition; geometric bases of computer science.
• **Tutorials:** ICCV'09 and CVPR'10 tutorials on sparse coding and dictionary learning for image analysis.

Funding ID
At UIUC I was PI for grants totalling about $2.8M in funding from the National Science Foundation and industry. Since returning to France, I have obtained the following awards:
• DGA (2008–2010): 2ACI. With Bertin, INRIA Rennes and Université de Caen. 130KEuros.
• DGA (2008): Itisecure. With E-vitech. 60KEuros.
• ANR: HFIMBR (2008–2010). With LASMEA and INRIA Rhône-Alpes. 110KEuros.
• ANR: Triangles (2008–2010). With INRIA Sophia-Antipolis and Lyon University. 5KEuros.
• MSR-INRIA lab (2008–2010): Image and video mining for science and humanities. 226KEuros.
• ANR-JST collaborative effort (pending): Image and video understanding for cultural heritage preservation. With INRIA Rhône-Alpes, University of Tokyo, and Keio University. 200KEuros.
• ANR (pending): Large-scale video access and understanding. With INRIA Rhône-Alpes, MRIM, INA, and EXALEAD. 200KEuros.

1c. 10-year track record

Top 10 publications as senior author:
For the 2000-to-present period alone, *Publish or Perish* records an h-number of **23** and over **3700** citations for my publications. My 10 most cited papers since 2000 are, according to Google Scholar:
1. S. Lazebnik, C. Schmid, J. Ponce, *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*, Proc. CVPR, II:2169-2178, 2006. **540** citations.

2. S. Lazebnik, C. Schmid, J. Ponce, *A Sparse Texture Representation Using Local Affine Regions*, PAMI, 27(8):1265-1278, 2005. **244** citations, including 106 for the CVPR'03 conference version.

3. F. Rothganger, S. Lazebnik, C. Schmid, J. Ponce, *3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints*, IJCV, 66(3):231-259, 2006. **198** citations, including 95 citations for the CVPR'03 conference version.

4. S. Lazebnik, C. Schmid, J. Ponce, *Affine-Invariant Local Descriptors and Neighborhood Statistics for Texture Recognition*, Proc. ICCV, 649-655, 2003. **76** citations.

5. S. Lazebnik, C. Schmid, J. Ponce, *Semi-Local Affine Parts for Object Recognition*, Proc BMVC, II:959-968, 2004. **71** citations.

6. Y. Furukawa and J. Ponce, *Accurate, Dense, and Robust Multi-View Stereopsis*, PAMI, 2010. In press. **60** citations for the CVPR'07 conference version.

7. S. Mahamud, M. Hebert, Y. Omori and J. Ponce, *Provably-Convergent Iterative Methods for Projective Structure from Motion*, Proc. CVPR, I:1018–1025, 2001. **54 citations.**

8. S. Lazebnik, C. Schmid, J. Ponce, *A maximum entropy framework for part-based texture and object recognition*, Proc. ICCV, I:832-838, 2005. **47 citations.**

9. Y. Furukawa and J. Ponce, *Carved Visual Hulls for Image-Based Modeling*, IJCV, 81(1):53–67, 2009. Special issue dedicated to the best papers of ECCV'06. **43** citations.

10. F. Rothganger, S. Lazebnik, C. Schmid, J. Ponce, *Segmenting, Modeling, and Matching Video Clips Containing Multiple Moving Objects*, PAMI, 29(3):477–491, 2007. **43** citations.

Notes: (1) PAMI and IJCV are the top two journals in computer vision. CVPR, ECCV, and ICCV are the most selective, peer-reviewed international conferences in computer vision, with lower acceptance rates than top journals, and papers often more cited than the subsequent journal publications. (2) I have a policy of appearing as last (senior) author on all papers with PhD students and junior colleagues. I have made significant contributions to all publications listed above. Their first author is one of my PhD students in all cases except [7], where he is a PhD student of my colleague M. Hebert.

Book:
• D. Forsyth and J. Ponce, **Computer Vision: A Modern Approach**, Prentice-Hall, 2003. This textbook has sold over 10,000 copies, and Chinese, Japanese, and Russian translations are available. **1801** citations.

Research monographs and chapters in collective volumes:

• C. Schmid, G. Dorko, S. Lazebnik, K. Mikolajczyk, J. Ponce, *Pattern Recognition with Local Invariant Features*, in **Handbook of Pattern Recognition and Computer Vision**, C.H. Chen and P.S.P Wang (eds.), World Scientific Publishing Co., 2004.

• J. Ponce, M. Hebert, C. Schmid, and A. Zisserman (eds.), **Toward Category-Level Object Recognition**, Springer-Verlag, Lecture Notes in Computer Science, Vol. 4170, 2007. **34** citations.

• S. Lazebnik, C. Schmid, J. Ponce, *Spatial Pyramid Matching*, in **Object Categorization: Computer and Human Vision Perspectives**, S. Dickinson (ed.), Cambridge University Press, 2009.

Granted patent:
• "Automated Reconfigurable Object Manipulation Device with an Array of Pins", S. Akella, S. Blind, C. Mc Cullough, and J. Ponce, US Patent # 6,633,797 (2003).

Invited presentations:
• *Keynote speaker:* Reconnaissance des Formes et Intelligence Artificielle, Toulouse (2004); European Workshop on Computational Geometry, Paris (2008); British Machine Vision Conference, Aberystwyth, Wales (2010).
• *Invited speaker:* Learning Workshop, Snowbird (2001); International Symposium on Core Research for Evolutional Science, Technology, Tokyo (2003); Workshop on Generic Object Recognition and Categorization, Washington DC (2004); Mathematics and Image Analysis Conference, Paris (2006); Rensselaer Polytechnic Institute CS Day (2006); ACCV'07 Vision Workshop, Hiroshima (2007); 3D Vision Workshop, Rio de Janeiro (2007); International Workshop on Shape Perception in Human and Computer Vision, Marseille (2008); International Workshop on Computer Vision, Venice (2008); Microsoft TechDays, Paris (2008); Beckman Institute 20th Anniversary Symposium, Urbana (2009); Conference on What Can Computer Vision Do for Neuroscience and Vice Versa, Janelia Farm, VA (2010).

Research expeditions: NA.

Organisation of international conferences:
• General chair, CVPR, Hilton Head Island, SC (2000).
• General chair, ECCV, Marseille (2008).
• Chair, International Workshop on Object Recognition, Taormina (2003, 2004, 2006).

International prizes/awards/academy memberships:
• IEEE Fellow (2003).

Membership to editorial board of international journals:
• Area editor, Computer Vision and Image Understanding (1994-2000).
• Associate editor, IEEE Transactions on Robotics and Automation (1998-2001).
• Editorial board member, International Journal of Computer Vision (2001–). (This includes serving as editor-in-chief from 2003 to 2007.)
• Editorial board member, Foundation and Trends in Computer Graphics and Vision (2005–).
• Editorial board member, SIAM Journal on Imaging Sciences (2009–).

1d. Extended synopsis of the project proposal

Digital video is everywhere, at home, at work, and on the Internet. Yet, effective technology for organizing, retrieving, improving, and editing its content is nowhere to be found. Models for video content, interpretation and manipulation inherited from still imagery are obsolete, and new ones must be invented. With a new convergence between computer vision, machine learning, and signal processing, the time is right for such an endeavor. Concretely, we will develop novel spatio-temporal models of video content learned from training data and capturing both the local appearance and nonrigid motion of the elements—persons and their surroundings—that make up a dynamic scene. We will also develop formal models of the video interpretation process that leave behind the architectures inherited from the world of still images to capture the complex interactions between these elements, yet can be learned effectively despite the sparse annotations typical of video understanding scenarios. Finally, we will propose a unified model for video restoration and editing that builds on recent advances in sparse coding and dictionary learning, and will allow for unprecedented control of the video stream. This project addresses fundamental research issues, but its results are expected to serve as a basis for groundbreaking technological advances for applications as varied as film post-production, video archival, and smart camera phones.

A video world. An enormous amount of resources is dedicated to the **creation** of digital video content (home movies, films, surveillance tapes, TV, video games), its **storage** (DVD libraries, DVRs, news archives), and its **distribution** (Institut National de l'Audiovisuel (INA), video on demand, YouTube). Effective general-purpose technology for **doing** something with this content in an automated or semi-automated fashion, on the other hand, is cruelly missing. I will focus in this proposal on two fundamental tasks:

1. **Understanding video content**—this is automatically answering queries such as **what** is happening in a scene, **who** is in it, **where** it is shot, and **when** some particular type of action is occurring.

2. **Manipulating video content**—that is **restoring** (e.g., deblocking, deblurring, denoising) damaged videos, and/or **editing** their content (e.g., adding, removing, replacing, or resizing objects and people).

Any solutions to these two problems will have an immediate impact on everyday life (e.g., organizing your family vacation clips), and applications in domains as varied as video archival (e.g., content-based retrieval), sociology (e.g., studies of cigarette use in sitcoms), entertainment (e.g., rig removal during post-production), or the camera phone industry (e.g., video denoising, crucial in this context with small lenses and sensing areas).

Of course, specialized tools have been developed for particular instances of these two problems, for example in surveillance [13] and sports [25] applications (Figure 1, top). As argued below, these do not generalize to unconstrained, real-life imagery as found in home videos, feature films, newscasts, and TV series. Likewise, existing technology for the specific problems of film restoration or rig and wire removal (Figure 1, bottom) typically require painstaking human intervention. Progress on both fronts requires scientific breakthroughs in computer vision, and will have immediate economic and societal impact. Thus, the first tenet of this proposal is that:

> **Developing a general framework for the understanding and manipulation of unconstrained video is today's frontier for computer vision research.**

This is the topic of this proposal.

Figure 1: **Top:** Sample frames from the Weizmann dataset (`http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`), a KTH football video (`http://www.nada.kth.se/cvap/res_widescreen_track.htm`), and French news footage from the 1950s, courtesy of INA. **Bottom:** Damaged frames in a film restoration scenario, with scratches and dirt (images courtesy of Laurent Joyeux and Louis Laborelli, INA), and an example of rig removal (the traffic light), reprinted from "The importance of invisible effects" [Wright'09].

The need to move forward. Human activity recognition in **restricted settings** has been the subject of active research for over 20 years. Much of it focuses on surveillance scenarios with little clutter and a fixed viewpoint ([13], Figure 1, top left), or stylized settings like sport events ([25], Figure 1, top center). Both are of course still very relevant today, but new methods are needed for **more general** scenarios involving home videos, feature films, newscasts, and TV series, with all the clutter, occlusions, personal interactions, and camera motion they entail. A slow shift toward this more realistic setting has taken place in the past couple of years, but today's methods rely on precise annotations (obtained manually or by automatically aligning subtitles and scripts [10]), implicitly assume fixed viewpoints and static cameras, and cannot handle hours of newscast or family footage with very sparse (if any) annotations. Understanding human activities and other "semantic" content is essential for effectively accessing, archiving and indexing video data. The ability to "intelligently" manipulate the content of a video is just as essential in many applications: This ranges from restoring old films (Figure 1, bottom left) or removing unwanted wires and rigs from new ones in post production (Figure 1, bottom right), to cleaning up a shot of your daughter at her birthday party, which is lovely but noisy and blurry because the lights were out when she blew the candles, or editing out a tourist from your Roman holiday video. Going beyond the modest abilities of current "digital zoom" (bicubic interpolation in general) so you can close in on that birthday cake, "deblock" a football game on TV, or turn your favorite DVD into a Blu-ray, is just as important.

A new alliance. I fear that a lack of communication between the computer vision and signal processing communities may have negative effects: computer vision researchers may wrongly perceive image processing techniques as a bit "old fashioned", while signal processing researchers may deplore the lack of comparative testing of computer vision algorithms against state-of-the-art image processing ones on established benchmark data. Thus, another theme of this project is a necessary convergence between computer vision, machine learning, and signal processing. The process has already started: For example, the idea of exploiting self-similarities in image analysis, originally introduced in computer vision for texture synthesis applications, is the basis for non-local means [4], one of today's most successful approaches to image restoration. In turn, by combining a powerful sparse coding approach

to non-local means [7] with modern machine learning techniques for dictionary learning [20], we have obtained denoising and demosaicking results that are the state of the art on standard benchmarks [22]. I will come back in much more detail to sparse coding and dictionary learning for image and video analysis in the second part of this proposal, but it is clear that a number of key algorithmic, mathematical, and more generally methodological tools are common to the computer vision, machine learning and signal processing communities. Others are likely to emerge in the next few years. Thus, another tenet of this proposal is that:

> **A new alliance between computer vision, machine learning, and signal processing is afoot, and this project is a unique opportunity to jump start it.**

The tools available today in each of these domains are not powerful enough, alone, to account for the diversity of content typical of everyday video. My third, and last tenet is thus that, as often in image analysis:

> **The main scientific challenges that stand in the way of true breakthroughs in video understanding and manipulation are modeling issues.**

This is argued below with three identified challenges.

Challenge 1: What is the right model for video content? In the world of still pictures, SIFT [19] has imposed itself as **the** local model of choice for image appearance in the past decade or so. Many variants have been proposed, but none does really better. This stems in part from the fact that SIFT is the result of years of careful design and experimentation with digital images, and incorporates the "know how" accumulated in the computer vision community for many years. In the dynamic world of video, no such consensus has emerged. Current local models like STIP [15] are fairly straightforward spatio-temporal extensions of SIFT and its variants, and they have not demonstrated their supremacy. At a more global level, human activities are depicted from a very large range of viewpoints in typical personal or professional video footage. Yet current approaches to action recognition use features that (implicitly) depend on viewpoint and/or assume static cameras. This is due once again, at least in part, to a relatively straightforward adaptation of models imported from the world of still images. Today, there is no satisfactory model for video content. Inventing the right one[3] is a major challenge, but also an opportunity to experiment with **new spatio-temporal models learned from training data, and capturing both the local appearance and the nonrigid motion of the elements that make up a dynamic scene**.

Challenge 2: What is the right model for video interpretation? Modern approaches to object recognition from still images are, by and large, minor variants of the so-called bag-of-features approach [6, 27], inherited from the text processing domain. There, it is reasonable to represent a document by the frequency vector (histogram) of the words that occur in it. There is of course no predefined dictionary of visual words, and various simple clustering methods have been used to construct such a vocabulary from SIFT features (for example) via vector quantization. Discarding all spatial information in the histogram of a bag of features builds some invariance to minor image transformations, but also throws away valuable spatial information. Adding some spatial structure to bags of features (in the form of spatial pyramids [17] for example) for related work) indeed improves recognition performance, but does not change the overall model structure. Recent efforts at activity recognition in unconstrained settings such as newscasts or TV series (as opposed to, say, surveillance scenarios, see [16, 23, 24]) are essentially straightforward spatio-temporal extensions of bags of features. This is not sufficient for capturing the complex interactions of the persons and objects present in typical

[3] Of course, I do not claim the existence of a single "right" model for video content, nor will I claim the existence of a single right model for the next two challenges. I use these words to emphasize the importance of the issue.

scenes. Thus, we will develop **formal models of the video interpretation process that capture these interactions, yet can be learned effectivey despite the sparse annotations typical of video understanding scenarios**.

Challenge 3: What is the right model for video manipulation? The question is worth asking for both the "low-level" (restoration) and "high-level" (editing) ends of the spectrum. Indeed, consumer-grade camcorders remain susceptible to noise at high sensitivity settings and/or low-light conditions, a problem that is exacerbated for camera phones with their small lenses and sensor areas. The classical problem of image and video restoration is thus still of acute and in fact growing importance. So are the related problems of deblurring, inpainting, and superresolution, that have received much attention lately with the emergence of computational photography. Tools for editing the content of a video—that is, for example, resizing, repainting, removing or adding objects and people in the scene it depicts, are also emerging. The underlying models vary greatly (epitomes, mosaics, layered representations, "soups" of patches, etc.), and a unified model for video restoration and editing is missing. Inventing such a model is the challenge of this part of the project. **Our recent work on sparse coding and dictionary learning for image understanding [20, 22] will form the backbone of this effort, with major extensions in several new directions, including encoding spatial consistency and task constraints in the learning, restoration, and editing tasks.**

Datasets and applications. A small, but significant part of this project will be dedicated to gathering new datasets. Existing ones are not sufficient because they are sometimes "too easy" (e.g., the Weizmann datasets) and perhaps sometimes "too hard" (e.g., the Hollywood datasets). They are not always representative of realistic tasks, but mix the effects of different factors, from occlusion and clutter, to interactions among people, or camera motion, which biases evaluation results. This part of our effort will be conducted in collaboration with M. Hebert at CMU. Although this project addresses fundamental research issues, its results are expected to serve as the basis for groundbreaking technological advances for applications as varied as film post-production, video archival, and smart camera phones. These will be explored in an opportunistic manner via our close contacts with end users, including F. Guichard at DXO, L. Laborelli and D. Teruggi at INA, P. Pérez at Technicolor, and S. Sullivan at Industrial, Light and Magic.

To conclude this part of the proposal, I will now try to assess the risks associated with this project.

High risk, high gain. Until the mid 1990s, object recognition from (still) images was limited to isolated, unoccluded objects on a uniform background or, equivalently to hand-segmented images; experiments were conducted on a handful of images, and typically restricted to specific objects ("is this **my** car?") instead of object categories (the much more difficult question "is this **a** car?"). Today, such category-level image classification is sometimes perceived (perhaps wrongly) as a solved problem, and very good results have also been obtained for the even more difficult detection problem ("**where** are the cars in this image?") on large and very challenging datasets such as those from the PASCAL VOC Challenge (`http://pascallin.ecs.soton.ac.uk/challenges/VOC/`). My goal with this project is to achieve a similar leap in performance for video understanding and manipulation. Here are two examples of specific success measures:

True scene understanding. What is in the scene? Where? What is happening to a person or in the scene? This goes further than generalizing to video today's still image technology. This goes back to one of the old and grand objectives of computer vision: scene understanding. This will be assessed on a large set of videos representative of movies, newscasts, as well as new datasets acquired during the course of the project.

True mid-scale inpainting. Today, we know how to replace a few missing pixels in an image using

its self-similarities (say erase text, such as sub-titles, for example), remove rigs and wires from a video by borrowing pixels from other frames with a cut-and-paste approach, or fill in very large areas of a photograph by borrowing content from another image that is overall similar. Despite attemps at respecting structural image constraints during the inpainting process, there is no satisfactory solution today to the problem of filling in mid-scale areas with complex internal structures, such as the missing wheel of a car for example. We will develop such a solution in this project.

Overall, this is a high-risk proposition: Can a small group achieve as much in five years as the computer vision community did in ten? But this is also a high-gain one.

Why there is a reasonable chance to succeed. This is a focused effort, with clear goals, and recent research by myself and other members of Willow is directly relevant to this proposal; see, for example (for the past two years only): [9, 16, 23] for action recognition and localization, [3] for models of the object recognition process, [22, 29] for image deblurring and restoration, and [20, 21] for sparse coding and dictionary learning for image classification. Several of these recent efforts have been conducted in close collaboration with machine learning (F. Bach and Y. LeCun) and signal processing (G. Sapiro) experts. This brings me to the next point.

The right team at the right time. After over 20 years in the US, I proposed upon my return to France in 2006 to create a computer vision team common to INRIA and Ecole Normale Supérieure (ENS). The Willow team officially started in the Spring of 2007. From the start, it was clear that machine learning was a key ingredient to new breakthroughs, and our activities have steadily grown in this area. In three short years, Willow has grown into a mature group of almost 30 people, and it divides its activities between computer vision, machine learning, and the cross-pollination of the two fields, with video as one of the core research areas. We have been very successful, with many publications in all the major international conferences and leading journals in both areas, but we are a large group with very diverse interests, ranging from camera geometry to statistics, and from image retrieval to bioinformatics applications of structured sparse coding, and I believe it is time to become lean again, by spinning off the core machine learning activities of Willow to a new group, headed by Francis Bach, who just received a Jr. ERC grant. The two teams will continue collaborating with each other (they will remain colocated at the INRIA site in central Paris), but they will have a sharper focus on their respective computer vision and machine learning activities.

The new, smaller Willow will consist of two permanent researchers besides myself—**Ivan Laptev** and **Josef Sivic**, whose main activities revolve around video understanding and image retrieval, plus several post-docs and about ten PhD students. The members of Willow form the core group for this project. A second circle of external collaborators complete the team: **Martial Hebert** (CMU) for computer vision, **Francis Bach** (INRIA) and **Yann LeCun** (NYU) for machine learning, and **Guillermo Sapiro** (Minnesota) for signal processing.

An opportunity for Europe. Historically, computer vision research has been associated with the US since its start in the 1960s. Its center of gravity started shifting toward Europe in the early 90s, with several groups gaining international recognition. The European Conference on Computer Vision is now considered to be in the same league as the top US and international conferences, CVPR and ICCV. When I chaired it in 2008, it brought together over 900 researchers (a 20% increase compared to the previous edition), with over 40% non Europeans and 200 American participants. At the same time, Asian countries, and China in particular, are emerging as strong competitors to Europe and the US. By recognizing the image/vision/learning convergence, and its incarnation through Willow and this project, Europe can gain a significant strategic advantage at the cutting edge of these fields, with breakthrough advances in video understanding and manipulation.

ERC Advanced Grant 2010
Research Proposal (Part B2)

2. The project proposal

We now present the state of the art and the overall objectives of our proposal, considering in turn its three main research challenges: modeling video, interpretation, and manipulation. We then present the main aspects of the proposed research, with five work packages, including the proposed formulation for the corresponding scientific problems and a research plan.

2a. State of the art and objectives

(i) Video content

Appearance models. In the world of still pictures, SIFT [19] has now imposed itself as *the* local model of choice for image appearance in matching and recognition taks. Many variants have been proposed, but none does really better. This stems in part from the fact that SIFT is the result of years of careful design and experimentation with digital images, and incorporates the "know how" accumulated in the computer vision community for many years. In the dynamic world of video, no such consensus has emerged. Current local models like STIP [15] are fairly straightforward spatio-temporal extensions of SIFT and its variants, and they have not demonstrated their supremacy. *Sparse coding* provides a promising alternative: Consider a signal (say an image patch or a spatio-temporal block of data) represented by a vector $\mathbf{x}$ in $\mathbb{R}^m$. We say that it admits a sparse approximation over a *dictionary* $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_p]$ in $\mathbb{R}^{m \times p}$ when one can find a linear combination of a "few" columns $\mathbf{d}_j$ of $\mathbf{D}$ that is "close" to the vector $\mathbf{x}$. Finding a sparse encoding of the signal $\mathbf{x}$ over $\mathbf{D}$ amounts to solving an optimization problem of the form

$$\boldsymbol{\alpha}^\star(\mathbf{x}, \mathbf{D}) = \arg\min_{\boldsymbol{\alpha}} ||\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}||_2^2 + \lambda\Phi(\boldsymbol{\alpha}),$$

where Φ is a sparsity-inducing regularizer, for example the ℓ_0 pseudo-norm counting the number of nonzero elements in $\boldsymbol{\alpha}$, the ℓ_1 norm, or the more robust *elastic-net* regularizer. Experiments have shown that modeling signals with such sparse decompositions (sparse coding) is very effective in many signal processing applications. For natural images, predefined dictionaries based on various types of wavelets have been used for this task. Introduced initially in [26] for modeling the spatial receptive fields of simple cells in the mammalian visual cortex, the idea of learning the dictionary from data instead of using off-the-shelf bases has been shown to significantly improve signal reconstruction. Learned dictionaries have also recently proven useful in image classification tasks.

Motion models. Much of the current work on human action representation and recognition from video sequences (implicitly) assumes that the action is observed from a fixed viewpoint [13, 25]. To achieve robustness to viewpoint changes, several recent methods exploit key frames observed from different viewpoints, geometric constraints (e.g., homographies induced by planar patterns, planar invariants, or rank constraints on motion parameters), or temporal self-similarities, but they usually assume that the camera remains static throughout filming. This is not realistic for many films and TV or homemade videos where the camera may undergo complex motions in a single shot (e.g., pan, tilt, and/or translation of a camera that may be hand-held or mounted on a tripod, a dolly, or a crane). This point is becoming crucial as the focus of today's research shifts from relatively simple videos as depicted in the Weizmann dataset for example to much more realistic ones such as those in the Hollywood datasets.

Objectives. Our goal in this part of the project is to develop new spatio-temporal models of video content that can effectively be learned from training data, and capture both the local appearance and the nonrigid motion of the elements that make up a dynamic scene.

(ii) Video interpretation

Template matching. Much of the work on activity recognition is aimed at surveillance scenarios with little clutter and fixed background, where silhouettes can be extracted reliably and used to define spatio-temporal templates [13], amenable to nearest-neighbor classification for example. Scenarios involving clutter and a moving camera are of course much more challenging.

Bags of features and their variants. Recent work has adapted to that setting the bag-of-features approach [6, 27] that has proven very successful in the still image domain: Briefly, local spatio-temporal features [15] are vector-quantized into predetermined visual words, and histograms of the words occurrences are classified using support vector machines (SVMs) for example [16]. In still image interpretation, plain bags of features, that discard all (global) spatial information, are often replaced by their structured variants, such as HOG descriptors or spatial pyramids, that retain some of this information. The same is true in action understanding [16, 23]. Likewise, SVMs may be replaced by other classifiers.

Weakly-supervised learning. Recent work on action recognition [16, 23] relies on textual annotations (obtained manually or through automated script alignment [10]) for learning action models. So will the approach to video understanding proposed in this project. It is important, however, to realize that textual annotations of video footage are often imprecise and quite sparse. On the other hand, unlabelled data is abundant. It is therefore desirable to develop semi- or weakly-supervised learning methods that can handle this type of data. Recent efforts in this direction, inlcuding ours [9] use the frameworks of discriminative clustering or multiple-instance learning.

Objectives. Our goal in this part of the project is to develop formal models of the video interpretation process that capture the complex interactions of the persons and object present in everyday videos, yet can be learned effectivey despite the sparse annotations typical of video understanding scenarios. We will adress both the classical problem of action recognition and detection, and that, more difficult, of scene understanding, with queries like: What is in the scene? Where? What is happening to a person or in the scene?

(iii) Video manipulation

Video restoration. With recent advances in sensor design, the quality of the signal output by digital SLRs and hybrid/bridge cameras is remarkably high. Point-and-shoot cameras, however, remain susceptible to noise at high sensitivity settings and/or low-light conditions, and this problem is exacerbated for mobile phone cameras with their small lenses and sensor areas, for both photographs and videos. Thus, the classical problem of image and video denoising is still of acute and in fact growing importance. So are the related problems of deblurring, demosaicking, and superresolution, that have received much attention lately with the emergence of computational photography (e.g., [12, 18]). Early work relied on various smoothness assumptions—such as anisotropic filtering, total variation, or image decompositions on fixed bases such as wavelets for example. More recent approaches include non-local means filtering, which exploits image self-similarities, learned sparse models, Gaussian scale mixtures, fields of experts, and block matching with 3D filtering (BM3D) [7]. As noted earlier, sparse coding using either predefined dictionaries or learned ones provides a very effective alternative in image restoration tasks [1, 22].

Video editing. Tools for editing the content of a video—that is, for example, resizing, repainting, removing or adding objects and people in the scene it depicts, are also emerging. The underlying models vary greatly (e.g., epitomes, mosaics, or layered representations [14]), and a unified model for video restoration and editing is missing. A popular approach to inpainting uses a "cut-and-paste" paradigm, by borrowing pixels from similar regions to fill in small holes in images, or from other frames to remove wires and rigs from a video, or by filling in very large areas of a photograph by borrowing content from another image that is overall similar. Despite attemps at respecting structural image constraints during the inpainting process (e.g., [5]), there is no truly satisfactory solution today to the problem of filling in mid-scale areas with complex internal structures.

Objectives. Our goal in this part of the project is to develop a unified model for video restoration and editing. Our recent work on sparse coding and dictionary learning for image understanding [20, 22] will form the backbone of this effort, with major extensions in several new directions, including encoding spatial consistency and task constraints in the learning, restoration, and editing tasks.

2b. Methodology

The project is structured into five work packages (WPs). The first four are methodological, and the last one is experimental.

WP1: Supervised sparse coding models of video content.
This WP proposes a general formulation for a supervised, task-oriented approach to sparse coding and dictionary learning. This formulation will be used to model the spatio-temporal appearance of video elements in WPs 3 and 4.
WP2: Motion models of video content.
This WP proposes to break away from view-dependent models of video content with a temporally local but spatially global model of nonrigid motion that is fully independent of camera motion. This model will be used in WP2 to gain a better understanding of the role of geometry and motion in the analysis of video footage, and as a feature in WP3 for video interpretation.
WP3: Modeling video interpretation.
This WP unifies several current models of visual recognition and generalizes them to the video setting using the models of video content developed in WP1 and WP2. The problems of weakly-supervised training, action recognition, and scene understanding will be addressed.
WP4: Modeling video manipulation.
This WP proposes a unified model of video restoration and editing. It generalizes the supervised sparse coding model of WP1 to take into account physical constraints associated with video restoration and spatial consistency constraints associated with video manipulation. The problems of deblurring, superresolution, and inpainting will be addressed.
WP5: Datasets.
This WP is dedicated to the construction of a new action dataset for controlled experiments in support of WPs 2 and 3.

In the upcoming sections, we will present in some detail, for each one of the methodological work packages (WPs 1 to 4), our initial formulation of the corresponding scientific problems. This is to ground our project in specifics, and also because these **all** correspond to novel, as of yet unpublished work. It is of course more than likely that, in a five-year project as ambitious as this one, they will evolve, and we also present detailed research plans for each WP.

WP1: Supervised sparse coding models of video content

1.1 Background

Learning local models of (spatio-temporal) appearance instead of using off-the-shelf ones is an attractive idea, and as discussed in Section 2a, this has led to significant progress in image and video restoration tasks [1, 22]. This is not surprising since these models are by construction adapted to reconstruction tasks. We now argue that sparse representations learned in a supervised, task-oriented manner can serve as a unified model of local appearance and lead to significant progress in video understanding and manipulation tasks.

1.2 Proposed formulation

Unsupervised sparse coding. Let us start with the simpler unsupervised case where a dictionary adapted to image reconstruction/restoration is learned from natural images or videos. Given a dictio-

Figure 2: Learning to classify image pixels as belonging to a bicycle or to the background [21]. This is an instance of sparse coding for image understanding where a dictionary is adapted to the task of discriminating between two object classes, here small images patches belonging to bicycles, and patches belonging to the background.

nary $\mathbf{D}$ in $\mathbb{R}^{m\times p}$, some signal $\mathbf{x}$ in $\mathcal{X}\subset\mathbb{R}^m$, and some code $\boldsymbol{\alpha}$ in $\mathbb{R}^p$, we can measure the discrepancy between $\mathbf{x}$ and its encoding by $\boldsymbol{\alpha}$ with the *elastic net* cost function

$$\varphi(\mathbf{x},\mathbf{D},\boldsymbol{\alpha}) \triangleq \frac{1}{2}||\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}||_2^2+\lambda_1||\boldsymbol{\alpha}||_1+\frac{\lambda_2}{2}||\boldsymbol{\alpha}||_2^2,$$

which is a sparsity-inducing regularizer in the sense defined in Section 2a. The corresponding sparse code $\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D})$ and residual function $\varphi^\star(\mathbf{x},\mathbf{D})$ for $\mathbf{x}$ given $\mathbf{D}$ are respectively defined by

$$\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D}) \triangleq \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\varphi(\mathbf{x},\mathbf{D},\alpha) \quad \text{and} \quad \varphi^\star(\mathbf{x},\mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\varphi(\mathbf{x},\mathbf{D},\boldsymbol{\alpha})=\varphi(\mathbf{x},\mathbf{D},\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D})), \tag{1}$$

and they are both uniquely defined when $\lambda_2>0$. In this context, traditional (unsupervised) dictionary learning amounts to finding the dictionary $\mathbf{D}$ in some convex subset $\mathcal{D}$ of $\mathbb{R}^{m\times p}$ that minimizes the *empirical cost*:

$$\min_{\mathbf{D}\in\mathcal{D}}\sum_{i=1}^n\varphi^\star(\mathbf{x}_i,\mathbf{D})=\min_{\mathbf{D}\in\mathcal{D},\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n\in\mathbb{R}^p}\sum_{i=1}^n\varphi(\mathbf{x}_i,\mathbf{D},\boldsymbol{\alpha}_i) \tag{2}$$

over n data points $\mathbf{x}_i$ $(i=1,\ldots,n)$.[4] However, as pointed out in [2], one is usually not interested in a perfect minimization of the empirical cost, but instead in the minimization of the *expected* cost—that is,

$$\min_{\mathbf{D}\in\mathcal{D}}\mathbb{E}_{\mathbf{x}}[\varphi^\star(\mathbf{x},\mathbf{D})] \tag{3}$$

where the expectation is taken relative to the (unknown) probability distribution $p(\mathbf{x})$ of the data, and is supposed to be finite. We have recently proposed [20] a very efficient online algorithm that iteratively solves the unsupervised dictionary learning problem by minimizing at each step a quadratic surrogate function of the empirical cost, and is guaranteed to converge to a stationary point of the expected cost function.

Supervised sparse coding. Once the dictionary $\mathbf{D}$ has been learned, the code $\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D})$ associated with each data vector $\mathbf{x}$ can be used as a feature vector in classification tasks for example. It is, however, preferable to adapt the dictionary to the task at hand. Previous approaches to this problem, including ours (Figure 2, see [21]) had to relie on heuristics to solve it in the specific case of classification tasks. We propose here a novel and general formulation for learning a dictionary in a *supervised* way for prediction tasks such as regression or classification. Concretely, suppose that we want to predict a variable $\mathbf{y}$ in $\mathcal{Y}$ from the observation $\mathbf{x}$, where $\mathcal{Y}$ is either an element of $\{0,1\}^q$ (or equivalently, a finite set of labels) in a classification task, or $\mathbb{R}^q$ in a regression task. We propose to jointly learn the dictionary $\mathbf{D}$ and a parameter matrix $\mathbf{W}$ for the task by solving the following optimization problem

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{W}\in\mathcal{W}} S(\mathbf{D}) \triangleq \mathbb{E}_{(\mathbf{y},\mathbf{x})}[\psi^\star(\mathbf{y},\mathbf{x},\mathbf{W},\mathbf{D})]+\frac{\mu}{2}||\mathbf{W}||_{\mathrm{F}}^2, \tag{4}$$

[4] In practice, one often takes $\mathcal{D}\triangleq\{\mathbf{D}\in\mathbb{R}^{m\times p}\}$ s.t. $\forall j\in\{1,\ldots,p\}$, $||\mathbf{d}_j||_2\leq 1$.

where μ is a regularization parameter, and $\psi^\star$ is a cost function adapted to the task, for example

$$\psi^\star(\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{D}) = \frac{1}{2}||\mathbf{y} - \mathbf{W}\boldsymbol{\alpha}^\star(\mathbf{x}, \mathbf{D})||_2^2, \tag{5}$$

where the elements of $\mathbf{y}$ are real for regression, and binary for classification. The same setting applies to loss functions other than the square loss, e.g., logistic regression.

The main difficulty of this optimization problem comes from its dependency on the solution of the nonsmooth optimization problem (1). Indeed, $\boldsymbol{\alpha}^\star$ is not differentiable with respect to $\mathbf{D}$. We have very recently proven, however, that, under mild assumptions (compact support for and continuity of p, $\mathcal{C}^2$ continuity of $\psi^\star$), the expected supervised cost S in Eq. (4) *is* differentiable, and that its gradient can itself be written "in closed form" as the expectation of a simple function of $\mathbf{W}$ and $\mathbf{D}$. In turn, it follows that the optimization of S is amenable to efficient projected first-order stochastic gradient techniques. Of course, this learning problem is not convex, but unsupervised sparse coding can be used to provide a reasonable initial estimate (although non-convex as well, unsupervised dictionary learning is empirically very stable under changes in initial conditions [20]). It should also be noted that, at test time, the computation of $\mathbf{y}$ is a convex problem for all convex losses, including the square and logistic ones.

1.3 Research plan

This part of the project will implement and extend the proposed problem formulation along several lines: **(1)** We will design a stochastic gradient algorithm for supervised sparse coding guaranteed to converge to a stationary point of the expected cost function (as we did in [20] for the *much* simpler unsupervised case), and construct an efficient implementation of this algorithm. **(2)** We will extend the proposed framework to the semi-supervised setting where some (or most) of the data is unlabelled, which is particularly relevant to video scenarios. **(3)** As a proof of concept, we will demonstrate the application of supervised sparse coding to image classification and deblocking on standard datasets, comparing its performance to the state of the art. **(4)** We will then demonstrate its application to video classification and manipulation tasks, as further explained in the presentation of WPs 3 and 4.

WP2: Motion models of video content

2.1 Background

We propose in this part of the project to focus on the dependency of human activity representation on camera motion by addressing the following problem: Assuming that a nonrigid scene is observed by a camera undergoing some unknown motion, can we construct a model of the corresponding video that is independent of that motion and affords an effective method for matching videos taken by two cameras with different motions? This is reminiscent of several approaches that generalize rigid structure-from-motion constraints to the nonrigid case by representing the nonrigid scene structures by linear combinations of elementary shapes or trajectories, assuming parametric (polynomial) motion models. In contrast with these methods, we will assume that, over a short period of time, the scene is globally nonrigid but locally rigid, and concentrate on modeling the relative motions of its locally rigid elements independently of the camera motion.

2.2 Proposed formulation

Factoring away camera motion. Concretely, let us assume short videos so that the motions of all objects of interest are globally nonrigid, but locally rigid within some image region. Specifically, we will divide each frame into a relatively small number of blocks—say, 100 blocks defined on a 10×10 grid, and track point features within each block over a short period of time—say, 30 frames, or 1s (Figure 3, left).[5] This model can be thought of as spatially global but temporally local since it

[5] Although it may be impossible to reliably track points across thousands of frames, it is quite reasonable with today's technology to expect being able to reliably track dozens of points across 30 frames or so.

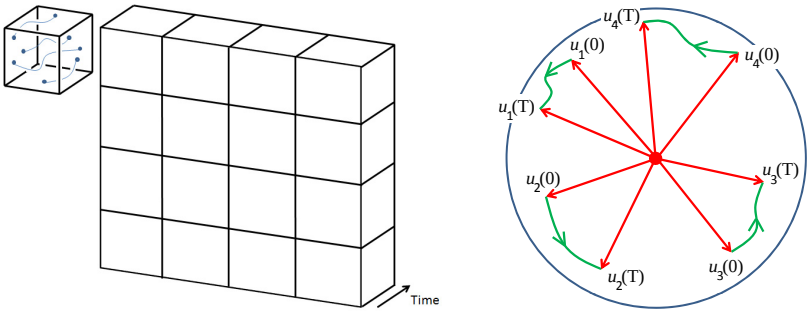

Figure 3: A spatio-temporal model for nonrigid motion. Left: Decomposition of a short clip into coarse spatio-temporal blocks, with the tracklets associated with one of the blocks shown in the upper left. Right: A spherical representation of the corresponding nonrigid motion: The rotations $\mathbf{P}_{st}(i)$ can be represented by the trajectories of the unit directions $\mathbf{u}_i(t)$ $(t = 0, \ldots, T)$ of the corresponding vectors $\mathbf{p}_{0t}(i)$ on the unit sphere and the magnitudes of these vectors along the trajectories. A similar representation can be used for the matrices $\mathbf{Q}_{ij}(t)$.

pools information from entire image frames into coarse blocks, but only uses a few frames at a time. Assuming a weak perspective projection model and rigid motion within each block, the corresponding "tracklets" can be fed to the Tomasi-Kanade factorization method and used to recover the motion of the corresponding rigid bodies at each frame up to an inherent ambiguity due to the mixture of camera and body motions. However, exploiting the fact that all the bodies are observed by the same (moving) camera, it is possible to reconstruct two sets of rotation matrices $\mathbf{P}_{st}(i)$ and $\mathbf{Q}_{ij}(t)$ that respectively describe the motion of a block i between two time frames s and t, and the relative orientation of two blocks i and j at a given time t, so that, if any other camera was filming the same scene while undergoing an arbitrary (and different) motion, the corresponding matrices would be related by:

$$\mathbf{P}'_{st}(i)\mathbf{R} = \mathbf{R}\mathbf{P}_{st}(i) \quad \text{and} \quad \mathbf{Q}'_{ij}(t)\mathbf{R} = \mathbf{R}\mathbf{Q}_{ij}(t) \tag{6}$$

for some rotation matrix $\mathbf{R}$ independent of i, j, s, and t. In other words, the matrices $\mathbf{P}_{st}(i)$ and $\mathbf{Q}_{ij}(t)$ provide a representation of the nonrigid motion of the scene elements that is only defined up to a (global) rotational ambiguity $\mathbf{R}$, but is *completely independent* of the camera motion.

A spherical model of nonrigid motion. It is well known that a rotation matrix $\mathbf{A}$ can be parameterized by a vector $\mathbf{a}$ such that $\mathbf{A} = \exp([\mathbf{a}]_\times])$. This vector is parallel to the axis of the rotation, with a magnitude equal to the rotation angle, and $[\mathbf{a}]_\times$ is the skew-symmetric cross-product operator such that, for any vector $\mathbf{b}$, $[\mathbf{a}]_\times\mathbf{b} = \mathbf{a} \times \mathbf{b}$. This exponential representation of rotations provides a convenient way to parameterize the matrices $\mathbf{P}_{st}(i)$ and $\mathbf{Q}_{ij}(t)$. Indeed, it can be shown that Eq. (6) is equivalent to

$$\mathbf{p}'_{st}(i) = \mathbf{R}\mathbf{p}_{st}(i) \quad \text{and} \quad \mathbf{q}'_{ij}(t) = \mathbf{R}\mathbf{q}_{ij}(t), \tag{7}$$

where, as before, lowercase vectors (e.g., $\mathbf{p}_{st}(i)$) are associated with the corresponding rotation matrices (e.g., $\mathbf{P}_{st}(i) = \exp([\mathbf{p}_{st}(i)]_\times)$. In particular, our motion model now admits a spherical representation in terms of the *unit directions* of the vectors $\mathbf{p}_{st}(i)$ and $\mathbf{q}_{ij}(t)$ and their magnitudes (Figure 3, right), so that the representations of videos of the same scene filmed by two cameras with arbitrary and different motions are separated by a rotation of the sphere.

Matching two videos. Each video can be represented by a graph whose nodes are the corresponding spatio-temporal blocks, and edges link nodes within some spatial neighborhood. Matching two videos reduces to minimizing

$$\sum_{i,s,t} ||\mathbf{p}'_{st}(\tau(i)) - \mathbf{R}\mathbf{p}_{st}(i)||_2^2 + \sum_{i,j,t} ||\mathbf{q}'_{\tau(i)\tau(j)}(t) - \mathbf{R}\mathbf{q}_{ij}(t)||_2^2 \tag{8}$$

with respect to the assignment function τ and the rotation $\mathbf{R}$. Here, the first sum is computed over the nodes of the graph and some fixed time samples s, t chosen a priori, and the second sum is computed over all the edges (i, j) of the graph and fixed times samples t.[6] Note that, for a given function τ, finding the rotation $\mathbf{R}$ minimizing (8) and thus best aligning the spherical models is easily reduced to an eigenvalue problem using quaternions.

Several formulations of this matching problem are possible: The first one is reminiscent of classical schemes for matching spherical representations of *shape* (as opposed to *motion* in our case), where the sphere is discretized into an icosahedron (or a finer subdivision if necessary), and the set of rotations aligning two discretized spheres can be explored efficiently. The second one explicitly solves a graph matching problem: When using the first term of (8) only, the optimization can be done using block coordinate descent, alternating steps where quaternions are used to estimate $\mathbf{R}$ with steps where the assignment problem of determining τ can be solved in cubic time by the Hungarian algorithm. Initial assignment guesses can be obtained without estimating $\mathbf{R}$ by matching rotations based on their angles. Taking the second term into account is more difficult, but standard techniques apply, in theory at least: Given the rotation $\mathbf{R}$, (8) can be written as a quadratic form in $\mathbf{\Pi}$, the permutation matrix (equal to 1 if point i is assigned to point j, and zero otherwise). The standard technique for solving such a problem is to consider local linear approximations (e.g., conditional gradient or power methods). In order to make this a descent algorithm, it may be necessary to subtract a term of the form $\mathrm{tr}(\mathbf{\Pi}\,\mathbf{\Pi}^\top)$ to make the objective function concave. This leads again to an alternating minimization algorithm.

2.3 Research plan

We will implement and extend the proposed problem formulation along several lines: **(1)** We will implement the proposed nonrigid motion model for video content, which will require handling ambiguities neglected so far in our presentation (e.g., the fact that a direction $\mathbf{u}$ and an angle θ define the same rotation as $-\mathbf{u}$ and $2\pi - \theta$). **(2)** We will develop and implement an efficient matching algorithm for discretized spherical representations of nonrigid motion. **(3)** We will develop an efficient graph matching approach to the same problem. **(4)** As a proof of concept, we will first demonstrate this implementation on videos of the same scene captured by different moving cameras. **(5)** We will then develop coding and indexing techniques to effectively match and compare videos of different actors filmed by different cameras but performing the same activities despite the variability in their play. This will be integrated in the video understanding strategies presented in WP3.

WP3: Modeling video interpretation

3.1 Background

Modern approaches to visual object, scene, or activity recognition usually follow the classical supervised classification paradigm where, given some global feature extraction process, the feature vectors associated with a number of positive and negative training samples are used to train a classifier. Once trained, this classifier can be used to classify test images or videos using the corresponding feature vectors, or to detect an object or an activity using the sliding window paradigm. We have shown recently [3] that many of these approaches can be decomposed into three steps (Figure 4, left):

1. **Filtering:** Some local descriptor is extracted at key points or on a dense grid, using linear filtering or a nonlinear operator.

2. **Coding:** The local descriptors are encoded using hard or soft vector quantization, or sparse coding into sparse binary of real vectors (codes).

[6] This assumes that there is a bijection between the two graphs. In practice, since the scene will be filmed from different viewpoints, some nodes and edges may not match. This can be handled by adding dummy nodes to both graphs.

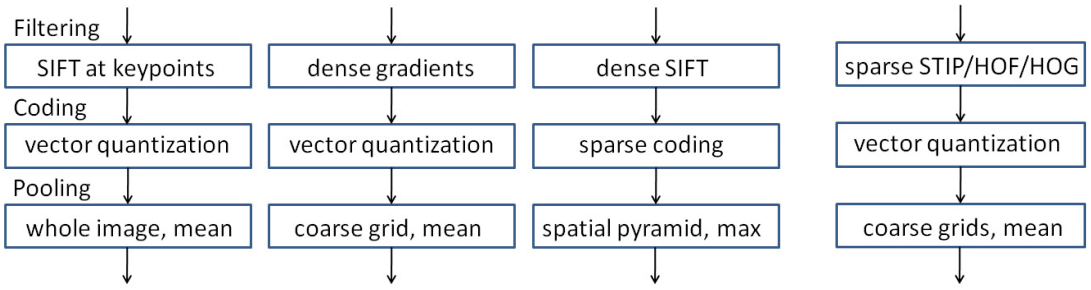

Figure 4: Left: Three approaches to (static) object recognition that fit the filtering/coding/pooling model: from left to right, bags of features [6], HOG [8], and a recent variant [30] of spatial pyramids. Right: The approach to action classification of [16] also naturally fits in this paradigm.

3. **Pooling:** A simple (elementwise mean or max) operator is applied to pool all codes in some spatial or spatio-temporal neighborhood into a single one and build robustness to image transformations and clutter.

The pooled codes associated with the entire image or spatio-temporal block are finally concatenated into a single feature vector, that can be passed to a classifier such as a support vector machine. Examples of this architecture include convolutional nets, bags of features, and spatial pyramids for example. Recent work on action recognition also fits in this framework (Figure 4, right).

3.2 Proposed formulation

Pooling and class separability. Consider a two-class categorization problem. Intuitively, classification is easier if the distributions from which points of the two classes are drawn have no overlap. In fact, if the distributions are simply shifted versions of one another (e.g., two Gaussian distributions with same variance), linear separability increases monotonically with the magnitude of the shift (e.g., with the distance between the means of two Gaussian distributions of same variance). We propose to examine how pooling affects the separability of the resulting feature distributions when the features being pooled are binary vectors (e.g., 1-of-K codes obtained by vector quantization in bag-of-features models). Our preliminary findings show that max pooling can sometimes increase, sometimes decrease the separation between the expected pooled feature values when compared to mean pooling, depending on the sample size P and the class-conditional probabilities of the feature being present at each individual sample. With mean pooling, the variance of the pooled feature decreases like $\frac{1}{P}$. Thus, it is always better to take into account all available samples of a given spatial pool in the computation of the mean. This is not the case with max pooling, where the variance can increase with pool size depending on the foreground distribution. We propose to conduct a thorough theoretical and empirical study of this class separability problem.

Discriminative sparse coding. As noted above, images and videos are often encoded before pooling into sparse binary or real vectors. In the simplest formulation of this approach [27], a codebook of K codewords is first learned by an unsupervised algorithm (e.g., K-means), and a binary, 1-of-K code $\boldsymbol{\alpha} \in \{0,1\}^K$ is obtained by minimizing the distance to the codebook: $\alpha_j = 1$ iff $j = \arg\min_{k \leq K} \|\mathbf{x} - \mathbf{d}_k\|_2^2$, where $\mathbf{d}_k$ denotes the k-th of the K codewords. The codes are then pooled over a (spatio-temporal) region of interest: In the case of mean pooling (histogramming) for example, this yields a feature vector of the form $\mathbf{h} = \frac{1}{P} \sum_{i \in \mathcal{N}} \boldsymbol{\alpha}_i$, where $\mathcal{N}$ denotes the region and P the number of features in that region. Van Gemert et al. [28] have improved this formulation by replacing hard by soft quantization. Instead of using a 1-of-K code where only the component corresponding to the nearest neighbor is non-zero, each codeword is assigned a score which reflects how closely it matches the input patch. This amounts to using codewords of the dictionary as centers of a Gaussian mixture model and performing coding as in the E-step of the expectation-maximization algorithm.

Figure 5: Sample frames of actions (left: "sit down", right: "open door") whose models have been learned in a weakly-supervised fashion, before being automatically detected in four movies [9].

Yang et al. [30] have shown that *unsupervised* (reconstructive) dictionary learning could be used to construct sparse codes $\boldsymbol{\alpha}$ for local image features, and demonstrated that this leads to significant classification performance improvement on classical benchmarks. Here, we propose instead to use a variant of our supervised sparse coding framework by optimizing a discriminative cost of the form

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{W}\in\mathcal{W}} \mathbb{E}_{\mathbf{y}}[||\mathbf{y} - \mathbb{E}_{\mathbf{x_y}}(\mathbf{W}\boldsymbol{\alpha}^\star(\mathbf{x_y},\mathbf{D}))||_2^2] + \frac{\mu}{2}||\mathbf{W}||_F^2, \tag{9}$$

where we assume mean pooling, and the signals $\mathbf{x_y}$ are part of the $\mathbf{y}$ class. As before, a logistic regression (or softmax for the multi-class case) loss could be used instead.

Beyond bags of features and their variants. As noted earlier, we plan to learn our models of video content from large amounts of automatically aligned videos and text. We will address both the classical action recognition problem and the more difficult one of scene understanding, where the goal is to parse a video in terms of its components (people and their surroundings) and their interactions. Thus, we will mine the text not only for action names, but for names of people, settings, and situations (see [23] for an early effort in this direction), with the goal of identifying objects, scenes, particular people, person attributes, human actions, and their relations/interactions. This will take advantage of modern technology for detecting and tracking people in videos. We believe that spatial (and temporal) reasoning will also prove crucial in this endeavor. In particular we plan to investigate spatio-temporal extensions of the deformable part model of Felzenszwalb et al. [11], and use the motion models constructed in WP2 to reason about the spatial relationship between scene components in localization/detection tasks.

Weakly-supervised learning. In typical video understanding scenarios, text annotations may be imprecise and quite sparse. Unlabelled data, on the other hand, is plentiful. The semi-supervised extension of our supervised sparse coding framework will prove particularly useful in this context. We also plan to use discriminative clustering methods to find and localize relevant instances of each concept in the video. We have already demonstrated that temporal human action detectors outperforming state-of-the-art methods can be learned using imprecisely aligned textual annotations from movie scripts (see [9] and Figure 5). We plan to continue this line of research and address spatial as well as temporal localization of human actions in the video.

3.3 Research plan

As with the other work packages, we will implement and extend the proposed problem formulation along several lines: **(1)** We will conduct a probabilistic theoretical investigation of the factors that influence different pooling strategies and validate it with experiments on synthetic and real data. **(2)** We will use our supervised approach to dictionary learning to implement discriminative sparse coding and validate this approach through comparisons with the state of the art on standard action recognition benchmarks. **(3)** We will investigate more complex models of visual recognition such as spatio-temporal extensions of the deformable part model of [11], and use the motion models constructed in WP2 as features to represent the spatial relationship between scene components in spatio-temporal localization tasks. **(4)** We will develop weakly- and semi-supervised methods for learning the corre-

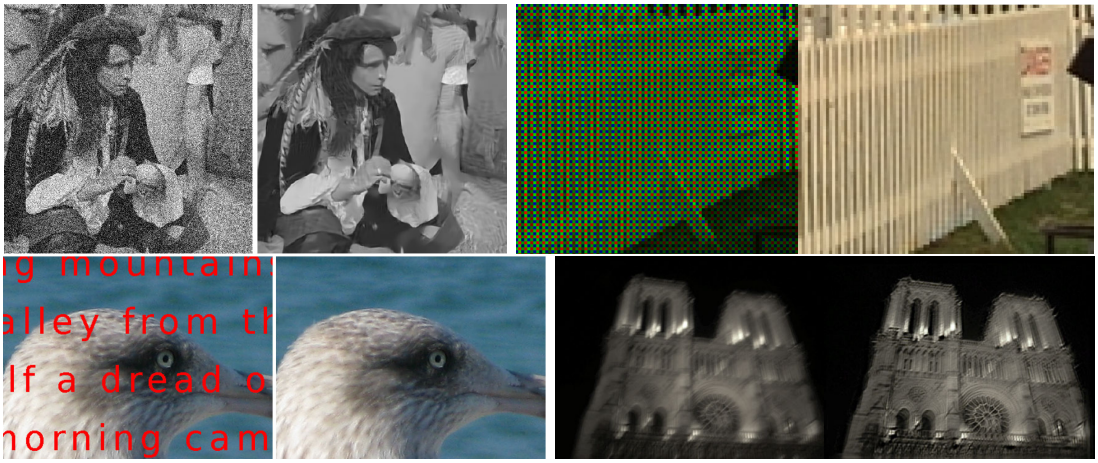

Figure 6: From left to right and top to bottom: examples of denoising (left: noisy image, right: restored one) [22], demosaicking (left: raw image with its Bayer pattern, right: reconstructed color image) [22], inpainting (left: original image, right: restored image with the text removed) [20], and deblurring (left: image with blur due to camera shake, right: deblurred image) [29]. Our denoising and demosaicking results are the state of the art on standard image processing benchmarks.

sponding video content models. **(5)** We will adapt our framework to video scene understanding, and evaluate its implementation on a large set of videos representative of movies and newscasts, as well as new datasets acquired during the course of the project.

WP4: Modeling video manipulation

4.1 Background

The *non-local means* approach [4] to image restoration exploits self-similarities in natural images to average out the noise among similar patches We have proposed in [22] to combine unsupervised dictionary learning with non-local means by using *simultaneous sparse coding* [7] to make similar patches share the same dictionary elements in their sparse decomposition. Experiments with images corrupted by synthetic and real noise have shown that this method outperforms the state of the art in both image denoising and image demosaicking tasks (Figure 6, top), making it possible to effectively restore raw images from digital cameras. In addition, we have recently demonstrated with our online unsupervised dictionary learning algorithm [20] that high-quality, small-scale inpainting (text removal) was possible for very large images (12MPixel, Figure 6, bottom left) at a reasonable cost. We are now poised to go much further by adapting these methods and our supervised sparse coding framework to the video domain.

4.2 Proposed formulation

Deblurring and superresolution. Let us assume a known, uniform blur, so the blurry image $\mathbf{B}$ is obtained from the sharp one $\mathbf{S}$ via convolution with some kernel $\mathbf{k}$—that is, $\mathbf{B} = \mathbf{k} \star \mathbf{S}$. For corresponding blurry and sharp patches $\mathbf{b}$ and $\mathbf{s}$ of these two images, we can define the error function

$$\psi^\star(\mathbf{s}, \mathbf{b}, \mathbf{W}, \mathbf{D}_s, \mathbf{D}_b) = \frac{1}{2}||\mathbf{s} - \mathbf{D}_b \boldsymbol{\alpha}^\star(\mathbf{b}, \mathbf{D}_b) - \mathbf{W}\mathbf{b}||_2^2,$$

and apply our task-driven dictionary approach to this problem by solving the optimization problem

$$\min_{\mathbf{D}_s, \mathbf{D}_b \in \mathcal{D}, \mathbf{W} \in \mathcal{W}} \mathbb{E}_{(\mathbf{s},\mathbf{b})}[\psi^\star(\mathbf{s}, \mathbf{b}, \mathbf{W}, \mathbf{D}_s, \mathbf{D}_b)] + \frac{\mu}{2}||\mathbf{W}||_{\mathrm{F}}^2. \tag{10}$$

At test time, the sharp image can be recovered, with $\mathbf{B}$, $\mathbf{D}_s$, $\mathbf{D}_b$ and $\mathbf{W}$ now fixed, by solving

$$\min_{\mathbf{S}} \sum_{i=1}^{n} \psi^\star(\mathbf{s}_i, \mathbf{b}_i, \mathbf{W}, \mathbf{D}_s, \mathbf{D}_b)] + \frac{\tau}{2}||\mathbf{k} \star \mathbf{S} - \mathbf{B}||_2^2. \tag{11}$$

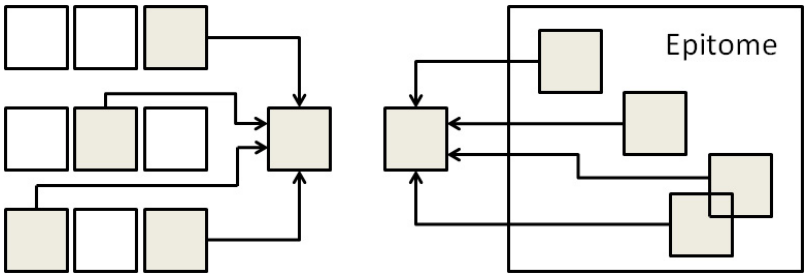

Figure 7: A flat dictionary (left) vs an epitome (right).

Note that Eqs. (10) for training and (11) for testing are consistent: Adding the term $\frac{\tau}{2}||\mathbf{k}\star\mathbf{S}-\mathbf{B}||_2^2$ to (10) does not change the result of this optimization problem. Adding the term $\frac{\mu}{2}||\mathbf{W}||_{\mathrm{F}}^2$ to (11) does not change its result either. This approach is easily extended to the superresolution problem.

Blind deblurring. The formulation of deblurring presented so far assumes, as is common in the image processing community [7], that the blur kernel is known. *Blind* deblurring, where the kernel must also be estimated, has also received quite a bit of interest in the last few years [18]. Our method can in principle be extended to this case when training pairs of sharp and blurry images associated with the same kernel $\mathbf{k}$ are available, $\mathbf{k}$ being learned at the same time as the other parameters of the model. We are also pursuing an alternative approach to blind removal of *non-uniform* blur under the assumption that it is due to camera shake, itself mostly accounted for by a rotation of the camera about its optical axis [29]. We plan to combine the two approaches into a single framework.

Epitomes. Jojic, Frey and Kanna [14] introduced in a probabilistic generative image model called an *epitome*. Intuitively, the epitome is a small image $\mathbf{E}$ that summarizes the content of a larger one, $\mathbf{X}$, and from which $\mathbf{X}$ can be reconstructed, denoised, etc. This is an intriguing notion, and epitomes have been extended to the video domain, where they have been used in denoising, superresolution, object removal and video interpolation. Aharon and Elad [1] have introduced an alternative formulation within the sparse coding framework. We want to go further in this project by proposing a notion of epitome adapted to our dictionary learning approach. We present below this formulation with image patches for simplicity, but it applies to spatio-temporal patches in a straightforward manner.

Sparse coding with an epitome is similar to sparse coding with a "flat" dictionary, except that the atoms are extracted from the epitome and may overlap instead of being chosen from an unstructured set of patches and assumed to be independent from each other (Figure 7). Concretely, let us denote by $\mathbf{E}$ the epitome of size q, and by p the number of (overlapping) patches of size m in $\mathbf{E}$.[7] Let us choose some arbitrary ordering for these patches, and denote by $\mathbf{R}_j$ the linear operator that extracts patch number j from the epitome ($j = 1, \ldots, p$). We define the linear operator $\mathbf{T} : \mathbb{R}^q \to \mathbb{R}^{m\times p}$ by

$$\mathbf{T}(\mathbf{E}) = [\mathbf{R}_1\mathbf{E}, \ldots, \mathbf{R}_p\mathbf{E}].$$

With this notation, we can define sparse coding just as before, by replacing the original function φ by

$$\varphi(\mathbf{x}, \mathbf{E}, \boldsymbol{\alpha}) \triangleq \frac{1}{2}||\mathbf{x} - \mathbf{T}(\mathbf{E})\boldsymbol{\alpha}||_2^2 + \lambda_1||\boldsymbol{\alpha}||_1 + \frac{\lambda_2}{2}||\boldsymbol{\alpha}||_2^2.$$

Likewise, epitome learning can be formulated as solving the optimization problem

$$\min_{\mathbf{E}\in\mathcal{E}} \mathbb{E}_{\mathbf{x}}[\varphi^\star(\mathbf{x}, \mathbf{E})], \tag{12}$$

where $\mathcal{E}$ is an appropriate convex domain for $\mathbf{E}$.

Inpainting with epitomes. Learned dictionaries can be used for small-scale inpainting (for removing text as in Figure 6, bottom left, for example). The task is much more difficult for larger areas: For

[7] We continue to identify image (and epitome) patches with vectors in $\mathbb{R}^m$.

example, although Criminisi et al. [5] try to preserve both texture and linear structure in their work, and achieve remarkable results (removing a person from a photograph for example), their approach is greedy, and there is no guarantee that structural details within the reconstructed area will be preserved. We propose to modify the epitomes proposed in the previous paragraph to enforce spatial consistency constraints, by requiring for example that overlapping patches use elements of the epitome that are within some preset distance from each other. It is also possible to enforce a global consistency constraint during inpainting. The corresponding optimization problem is of course more difficult in this case.

4.3 Research plan

We propose in this work package to implement and extend the proposed formulation. **(1)** We will implement the proposed approach to video deblurring and superresolution. **(2)** We will extend it to the blind deblurring case. **(3)** We will implement our sparse coding approach to epitome construction, and test it in inpainting tasks. **(4)** We will develop an efficient algorithm for imposing local spatial consistency constraints during epitome learning and global spatial consistency during testing. **(5)** We will use this algorithm for inpainting tasks with large missing areas and complex internal structures, such as the missing wheel of a car for example. In all cases, the developed video manipulation will first be tested on still images first as a proof of concept, then extended to videos and compared to the state of the art on standard benchmarks.

WP5: Datasets

Annotated data is always hard to come by in computer vision. This is particularly true in the video domain, where manual annotation is very time consuming. As others [16, 23], we will rely in this project on the rich (but temporally imprecise) annotations obtained by automatically aligning subtitles and scripts [10] in datasets such as the Hollywood ones [16, 23]. We will also rely on the much sparser annotations available in newscasts thanks to the work of archivists at the "Institut National de l'Audiovisuel" (INA). This data is available to us through an existing collaborative project with INA and the MSR-INRIA laboratory in Saclay. Existing datasets are sometimes "too easy" (e.g., the Weizmann datasets) and perhaps sometimes "too hard" (e.g., the Hollywood datasets): They are not always representative of realistic tasks, but mix the effects of different factors, from occlusion and clutter, to interactions among people, or camera motion, which biases evaluation results. Thus, we will create, manually annotate, and make publicly available an action dataset of our own. It will feature multiple, controlled scenarios, including combinations of **(1)** multiple fixed or mobile cameras to understand the dependency of action recognition on viewpoint and camera motion, **(2)** isolated actors performing the same action to understand its dependency on inter-actor variability, **(3)** multiple actors interacting with each other, and **(4)** simple and more complex backgrounds (see `http://www.cs.rochester.edu/~rmessing/uradl/` for a related effort with fixed viewpoint). This dataset will be annotated manually during its creation, and used to support controlled experiments for WPs 2 and 3. This part of our effort will be conducted in collaboration with M. Hebert at CMU, and will involve filming volunteers in our lab. We will seek the advice of ethics experts at INRIA and the Cap Digital competitiveness cluster before any use of this data.

References

[1] M. Aharon and M. Elad. Sparse and redundant modeling of image content using an image-signature dictionary. *SIAM Journal of Imaging Sciences*, 1(3):228–247, 2008.

[2] L. Bottou and O. Bousquet. The trade-offs of large scale learning. In *Proc. Neural Info. Proc. Systems*, pages 161–168, 2008.

[3] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2010. Accepted for publication.

[4] A. Buades, B. Coll, and J.M. Morel. A non-local algorithm for image denoising. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2005.

[5] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based inpainting. *IEEE Trans. on Image Processing*, 13(9):1200–1212, 2004.

[6] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[7] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Trans. on Image Processing*, 16(8):2080–2095, 2007.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, volume II, pages 886–893, 2005.

[9] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *Proc. Int. Conf. Comp. Vision*, 2009.

[10] M. Everingham, J. Sivic, and A. Zisserman. 'Hello! my name is... Buffy' - Automatic naming of characters in TV video. In *British Machine Vision Conference*, pages 889–908, 2006.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2008.

[12] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Proc. Int. Conf. Comp. Vision*, 2009.

[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Patt. Anal. Mach. Intell.*, 29(12):2247–2253, 2007.

[14] N. Jojic, B.J. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *Proc. Int. Conf. Comp. Vision*, volume I, pages 34–41, 2003.

[15] I. Laptev. On space-time interest points. *Int. J. of Comp. Vision*, 64(2/3):107–123, 2005.

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2008.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, volume II, pages 2169–2178, 2006.

[18] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2009.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Comp. Vision*, 60(4):91–110, 2004.

[20] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[21] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2008.

[22] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Proc. Int. Conf. Comp. Vision*, 2009.

[23] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2009.

[24] J.C. Niebles and H. Wang L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. of Comp. Vision*, 79(3), 2008.

[25] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking – Linking identities using Bayesian network inference. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2006.

[26] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[27] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. Int. Conf. Comp. Vision*, 2003.

[28] J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, and A.W.M. Smeulders. Kernel codebooks for scene categorization. In *Proc. European Conf. Comp. Vision*, 2008.

[29] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2010.

[30] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2009.

2c. Resources

(i) Core team members
The core team members are the PI and the members of the Willow team, which consists of two permanent researchers besides myself—**Ivan Laptev** and **Josef Sivic**, whose main activities revolve around video understanding and image retrieval, several post-docs and about ten PhD students.

(ii) External team members
A second circle of four external collaborators complete the team: **Martial Hebert** (CMU) for computer vision, **Francis Bach** (INRIA) and **Yann LeCun** (NYU) for machine learning, and **Guillermo Sapiro** (Minnesota) for signal processing.

To promote the development of an alliance between computer vision, machine learning, and signal processing, we will also organize regular workshops with a third circle of researchers, who although not directly involved in the project, will contribute their ideas and expertise: D.A. Forsyth (UIUC), F. Durand (MIT), A. Efros (UIUC), and C. Schmid and her team (INRIA) in computer vision; M. Elad (Technion), S. Mallat (Polytechnique), and J.-M. Morel (ENS Cachan) in signal processing.

Applications are not the main focus of this project, but we will pursue them in an opportunistic manner through our contacts in the industry: F. Guichard (DXO – camera phones), L. Laborelli and D. Teruggi (INA – video archival), P. Pérez (Technicolor – post production), S. Sullivan (ILM – post production).

(iii)Available resources Our current grants are listed below:
- DGA (2008–2010): 2ACI. With Bertin, INRIA Rennes and Université de Caen. 130KEuros.
- DGA (2008): Itisecure. With E-vitech. 60KEuros.
- ANR: HFIMBR (2008–2010). With LASMEA and INRIA Rhône-Alpes. 110KEuros.
- ANR: Triangles (2008–2010). With INRIA Sophia-Antipolis and Lyon University. 5KEuros.
- MSR-INRIA lab (2008–2010): Image and video mining for science and humanities. 226KEuros.

All of these grants, except for the MSR-INRIA one, which is expected to be renewed for three years in 2010, will have expired by the beginning of this project. Two more grants are pending:
• ANR-JST collaborative effort (pending): Image and video understanding for cultural heritage preservation. With INRIA Rhône-Alpes, University of Tokyo, and Keio University. 200KEuros.
• ANR (pending): Large-scale video access and understanding. With INRIA Rhône-Alpes, MRIM, INA, and EXALEAD. 200KEuros.

If funded, these grants will provide additional resources for the Willow activities in video analysis.

The Willow team is equipped with numerous PCs as well as a 128-core PC cluster.

(iv) Requested resources and project costs.

Personnel costs:
• 110KEuros per year for the PI's salary with a 70% commitment of his time.
• 43,4Keuros per year for 5 years to cover the travel and living expenses of our foreign academic partners.
• 100KEuros per year to fund 2 post-docs per year for 5 years. The post-doc salary is at a higher-than-usual level to allow us to hire the best possible post-docs: The best French PhD students should go abroad for their post-doc, and the usual French salaries are too low to attract top foreign post-docs.
• 70.4KEuros per year to fund 2 PhD students for years 1 to 3 of the project.
• 35.2KEuros per year to fund 1 PhD student for years 2 to 4 of the project.
• 70.4KEuros per year to fund 2 PhD students for years 3 to 5 of the project.

Other direct costs:
• 30KEuros of travel per year for 5 years to visit our partners and attend international conferences.
• 50KEuros during years 1 and 4 to upgrade our cluster.

The rest of the costs consists of eligible indirect costs, at the rate of 20% of the direct costs. The grand total amounts to 2,454,090Euros over a period of 5 years, as detailed below.

	Cost category	Year 1	Year 2	Year 3	Year 4	Year 5	Total
	Personnel						
Direct costs	Principal investigator	110,000	110,000	110,000	110,000	110,000	550,000
	Invited profs. (6 mths/yr)	43,445	43,445	43,445	43,445	43,445	217,225
	Post-docs (2 per yr)	100,000	100,000	100,000	100,000	100,000	500,000
	PhD student	35,190	35,190	35,190			105,570
	PhD student	35,190	35,190	35,190			105,570
	PhD student		35,190	35,190	35,190		105,570
	PhD student			35,190	35,190	35,190	105,570
	PhD student			35,190	35,190	35,190	105,570
	Total personnel:	323,825	359,015	429,395	359,015	323,825	1,795,075
	Other direct costs:						
	Equipment (cluster)	50,000			50,000		100,000
	Travel	30,000	30,000	30,000	30,000	30,000	150,000
	Total other direct costs	80,000	30,000	30,000	80,000	30,000	250,000
	Total direct costs	403,825	389,015	459,395	439,015	353,825	2,045,075
Indirect costs (overheads)	Max 20% of direct costs	80,765	77,803	91,879	87,803	70,765	409,015
Total costs of project		484,590	466,818	551,274	526,818	424,590	2,454,090
Requested grant		484,590	466,818	551,274	526,818	424,590	2,454,090

For the above cost table, please indicate the % of working time the PI dedicates to the project over the period of the grant	70%

2d. Ethical issues

Please see the Ethical Issues Annex for an explanation of the privacy ethical issues involved with creating video datasets, and how they will be dealt with appropriately.

Ethical issues Table:

	Research on Human Embryo/ Foetus	YES	NO
	Does the proposed research involve human Embryos?		NO
	Does the proposed research involve human Foetal Tissues/ Cells?		NO
	Does the proposed research involve human Embryonic Stem Cells (hESCs)?		NO
	Does the proposed research on human Embryonic Stem Cells involve cells in culture?		NO
	Does the proposed research on Human Embryonic Stem Cells involve the derivation of cells from Embryos?		NO
	DO ANY OF THE ABOVE ISSUES APPLY TO MY PROPOSAL?		NO

	Research on Humans	YES	NO
	Does the proposed research involve children?		NO
	Does the proposed research involve patients?		NO
	Does the proposed research involve persons not able to give consent?		NO
	Does the proposed research involve adult healthy volunteers?		NO
	Does the proposed research involve Human genetic material?		NO
	Does the proposed research involve Human biological samples?		NO
	Does the proposed research involve Human data collection?		NO
	DO ANY OF THE ABOVE ISSUES APPLY TO MY PROPOSAL?		NO

	Privacy	YES	NO
	Does the proposed research involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?		NO
	Does the proposed research involve tracking the location or observation of people?		YES
	DO ANY OF THE ABOVE ISSUES APPLY TO MY PROPOSAL?		YES

	Research on Animals	YES	NO
	Does the proposed research involve research on animals?		NO
	Are those animals transgenic small laboratory animals?		NO
	Are those animals transgenic farm animals?		NO
	Are those animals non-human primates?		NO
	Are those animals cloned farm animals?		NO
	DO ANY OF THE ABOVE ISSUES APPLY TO MY PROPOSAL?		NO

	Research Involving Developing Countries	YES	NO
	Does the proposed research involve the use of local resources (genetic, animal, plant, etc)?		NO
	Is the proposed research of benefit to local communities (e.g. capacity building, access to healthcare, education, etc)?		NO
	DO ANY OF THE ABOVE ISSUES APPLY TO MY PROPOSAL?		NO

	Dual Use	YES	NO
	Research having direct military use		NO
	Research having the potential for terrorist abuse		NO
	DO ANY OF THE ABOVE ISSUES APPLY TO MY PROPOSAL?		NO

Other Ethical Issues	YES	NO
Are there **OTHER** activities that may raise **Ethical Issues**?		NO
If **YES** please specify:		NO

3. Research environment

The PI and his research team, Willow, are part of the Laboratoire d'Informatique de l'Ecole Normale Supérieure (LIENS), a Joint INRIA/ENS/CNRS Research Unit (JRU). Two of the members of the LIENS are members of the French Academy of Sciences, one received the CNRS gold medal in 2007, one is a Turing price winner, two have received an ERC advanced grant, and one has received an ERC junior grant.

INRIA will be physically hosting the project in its Place d'Italie offices in the center of Paris.

3a. PI's host institution

INRIA, the French national institute for research in computer science and control, operating under the dual authority of the Ministry of Research and the Ministry of Industry, is dedicated to fundamental and applied research in information and communication science and technology (ICST). The Institute also plays a major role in technology transfer. The PI will be hosted at INRIA Paris-Rocquencourt. INRIA Paris-Rocquencourt is one of the eight research centers of INRIA and is composed of 34 research teams. It is located in Rocquencourt, West of Paris, with INRIA's head office, and an additional branch opened in central Paris in 2009. INRIA Paris-Rocquencourt's research teams take part in many European and international projects. A third of the permanent researchers recruited in the past few years comes from abroad (from Europe, Japan, United States of America, etc.). INRIA Paris-Rocquencourt also hosts 170 foreign researchers every year. It has established many joint ventures, and plays an important part in Cap Digital, Moveo and System@tic competitiveness clusters. INRIA Paris-Rocquencourt is a founding member of Digiteo advanced research thematic network. In close collaboration with Universities and higher education schools in the Ile-de-France area, INRIA Rocquencourt considers training by and for research to be particularly important. INRIA Paris-Rocquencourt takes part in the main Information and Communication Sciences and Technologies, Masters degree courses with currently about 150 PhD students.

It is worth noting that the organization of INRIA in "project-teams" is particularly well suited to the ERC Advanced Grant program: An INRIA project-team like Willow brings together a group of researchers, post-docs, and PhD students, under the leadership of an experienced scientist. Their common goal is to address a particular scientific and technological challenge. INRIA project-teams must be approved by an evaluation committee particularly qualified in the corresponding scientific field, and are thoroughly reviewed by a similar scientific committee every four years. Each INRIA project-team enjoys organisational and management autonomy. It decides what use to make of its own financial resources. It can also draw on the resources made available by the "research support" services of INRIA's eight regional centres (development and transfer, human resources, funding, information technology, communication, etc.). The Institute promotes exchange and collaboration among the project-teams of its eight centres. It also encourages them to cooperate with their counterparts in other countries.