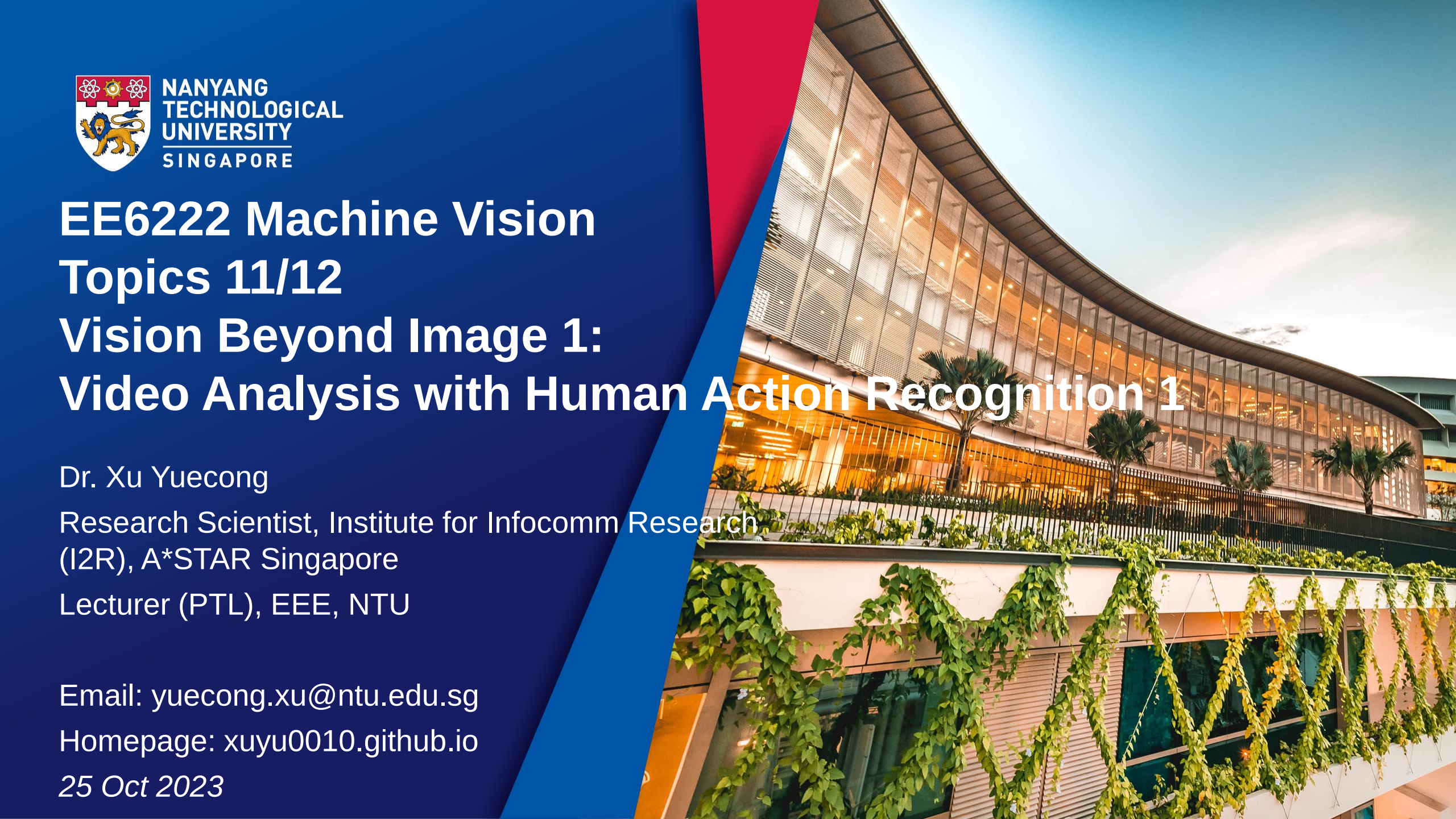


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

EE6222 Machine Vision
Topics 11/12
Vision Beyond Image 1: Video Analysis with Human Action Recognition 1

Dr. Xu Yuecong

Research Scientist, Institute for Infocomm Research (I2R), A*STAR Singapore

Lecturer (PTL), EEE, NTU

Email: yuecong.xu@ntu.edu.sg

Homepage: xuyu0010.github.io

25 Oct 2023

Self Introduction

- Research Scientist at Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore. Part-time Lecturer, NTU, EEE.
- B. Eng. 2017, NTU; Ph. D. 2021, NTU (Nanyang President's Graduate Scholarship).
- Research focus on RGB video analysis, action recognition, video domain adaptation, video captioning, multi-modal video analysis, time-series analysis.
- Multiple papers in video analysis, including ICCV (2021, 2023), ECCV (2022).
  - Xu, Y., Yang, J., Cao, H., Chen, Z., Li, Q., & Mao, K. (2021). Partial video domain adaptation with partial adversarial temporal attentive network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9332-9341).
  - Xu, Y., Yang, J., Cao, H., Wu, K., Wu, M., & Chen, Z. (2022). Source-free video domain adaptation by learning temporal consistency for action recognition. In European Conference on Computer Vision (pp. 147-164).
  - Xu, Y., Yang, J., Zhou, Y., Chen, Z., Wu, M., & Li, X. (2023). Augmenting and Aligning Snippets for Few-Shot Video Domain Adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9332-9341).
- (Co-)Organizer UG2+ Challenge held in conjunction with CVPR 2021/2022/2024.
- Reviewer of top conference and journals, including ICLR, ACM MM, TMM, TNNLS, TKDD.

Topics and References

Topics Outline:

- Characteristic of videos and video analysis. (11)
- Representation-based (Traditional) video feature extraction. (11)
- (Deep) Learning-based video feature extraction. (12)
- Latest developments in video analysis (mostly with deep learning) (12)

References:

- Textbooks recommended by Prof. Jiang are helpful in understanding video analysis.
- Overall understanding: survey papers on video analysis.
- Latest developments: papers from top conferences and top journals.

Example References

- Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5), 1366-1401.
- Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing*, 60, 4-21.
- Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, 28(6), 976-990.
- Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision* (ECCV) (pp. 803-818). (**TRN**)
- Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308). (**I3D**)
- Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In Proceedings of the *IEEE/CVF international conference on computer vision* (pp. 6202-6211). (**SlowFast**)
- Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In Proceedings of the *IEEE/CVF conference on computer vision and pattern recognition* (pp. 244-253). (**ViT**)
- Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3202-3211). (**Video Swin Transformer**)

Goal of Topics 11 – 12

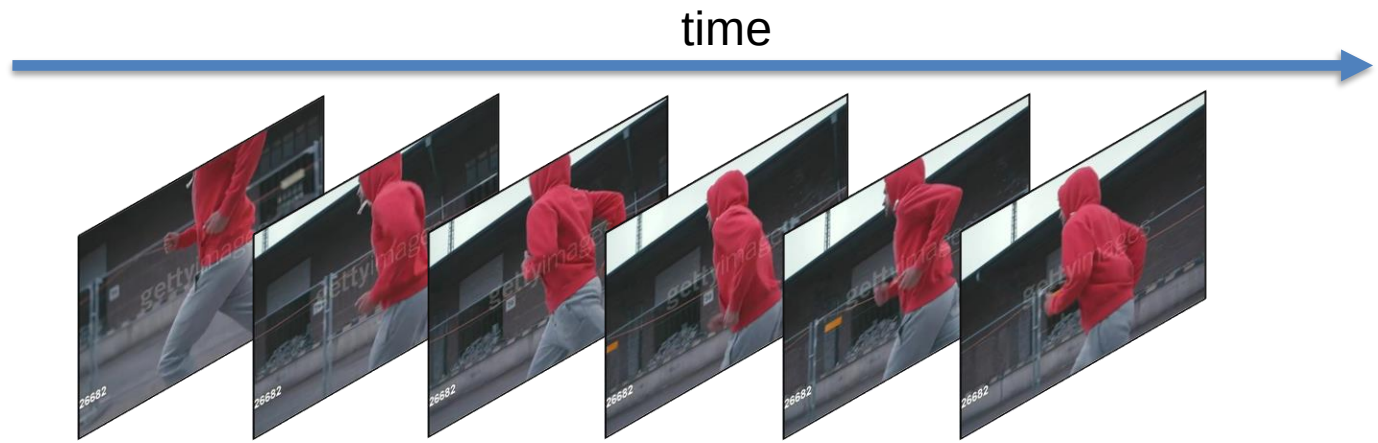
- ✓ Basic concepts in video analysis.
- ✓ Basic knowledge in the development of video analysis.
- ✓ Develop critical thinking in video analysis.
- ✓ Start from the basics, learn how to learn.

- ✓ **Conceptual** (not much computation).
- ✓ **Understand**, not recitation.

Continuous Assessment (CA) (15+10%)

- Purpose of Assessment: a simple project to attempt human action recognition (HAR) in an under explored scenario.
- Understand how to perform HAR intuitively (from scratch):
  - The workflow of a HAR research project.
  - Present and discuss the outcome.
- Points will be given for each step.
- Optional points will be given for further exploration.
- No example code will be given.
- Follow the additional directions clearly and precisely.
- Start early and ask questions early.
  - No dedicated office hours (yet), please email. Offline Q&A would be available upon appointment.
  - 3 attempts for submission.

Video Fundamentals: from Images to Videos

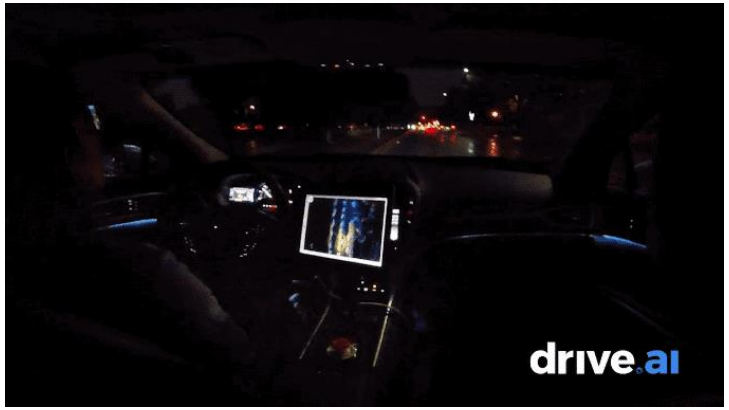


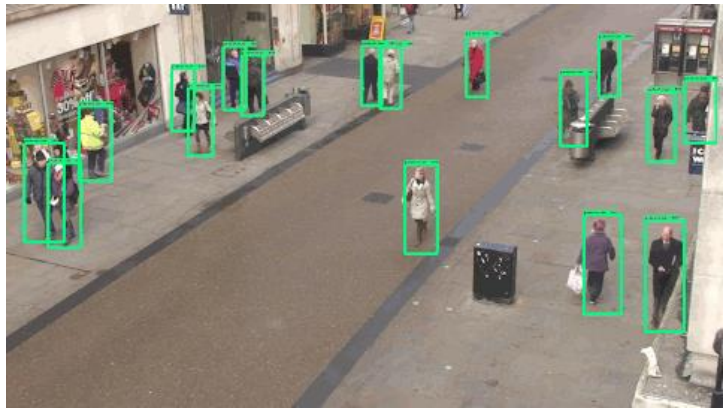
- A video can be viewed as a **sequence** of images.
  - By default, in computer vision the sequence of images are **closely related**.
- A video is more than a sequence of images:
  - Other modalities (types) of information are also included, e.g., audio.

Video Fundamentals: Video Analysis Real-world Applications

Security Surveillance

Smart Manufacturing

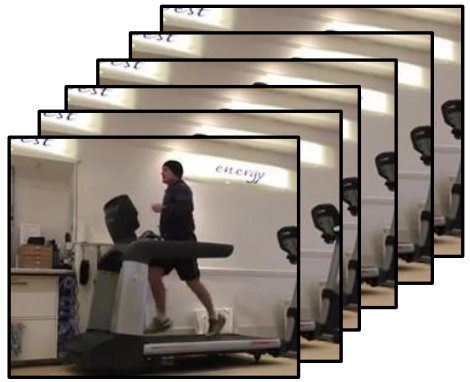




Autonomous Driving

Video Fundamentals: Characteristics of Videos

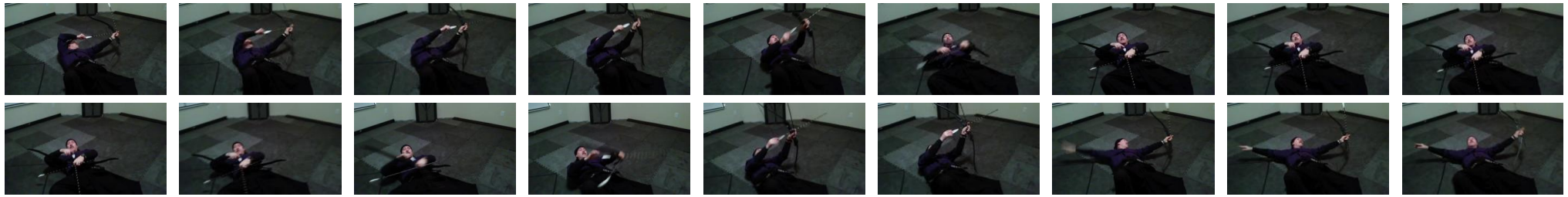
- Contains much more information than a single image:
  - Without considering information of different modalities (e.g., audio and flow-based)
  - Most videos are shot with 30 FPS (frames per second)
  - A 1-minute video usually contains 1800 images. (For 120 FPS videos, a 1-minute video contains 7200 images)
- Usually expressed as a 4D-Tensor (only consider RGB frames):
  - $T \times 3 \times H \times W$ or $3 \times T \times H \times W$ (3 for RGB channels)
  - How to start video analysis given the large amount of information?
  - What is the workflow of video analysis?



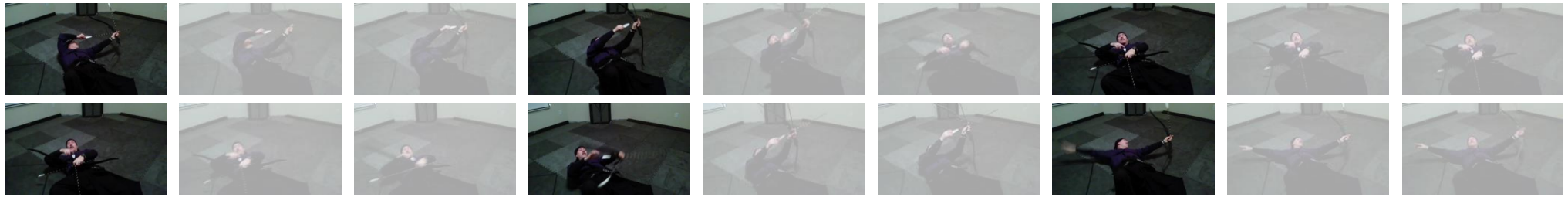
Video Fundamentals: Characteristics of Video Analysis

- Problem of using raw video for video analysis:
  - Videos are too big to process and analyze (for uncompressed video: ~GB per minute).
  - Solution: Frame Sampling (Uniform Sampling; Random Sampling)

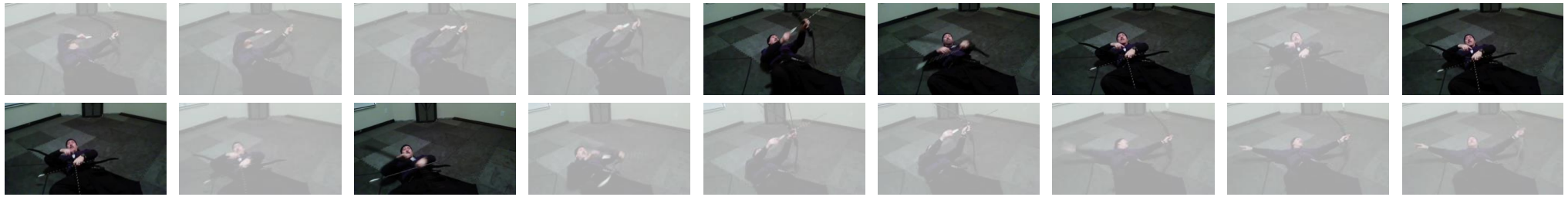
Original Video



Uniform Sampling



Random Sampling



Video Fundamentals: Video Sampling

- Different sampling methods could produce very different input.
- The impact of different sampling methods depends on **various factors**, e.g., the complexity of the video, the length of the video, the number of sampling frames.

Uniform Sampling

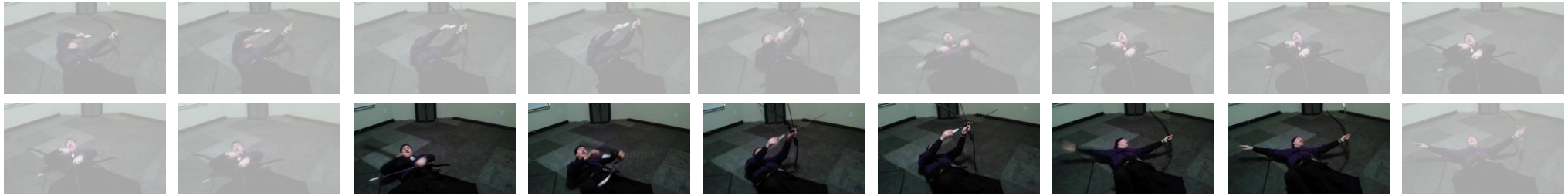
Random Sampling

- No obvious conclusion on which is better, highly depends on data.

Video Fundamentals: Video Sampling

- More recently: learning-based motion-guided sampling (MGSampler) (Zhi, Y., Tong, Z., Wang, L., & Wu, G. (2021). Mgsampler: An explainable sampling strategy for video action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1513-1522).

MGSampler



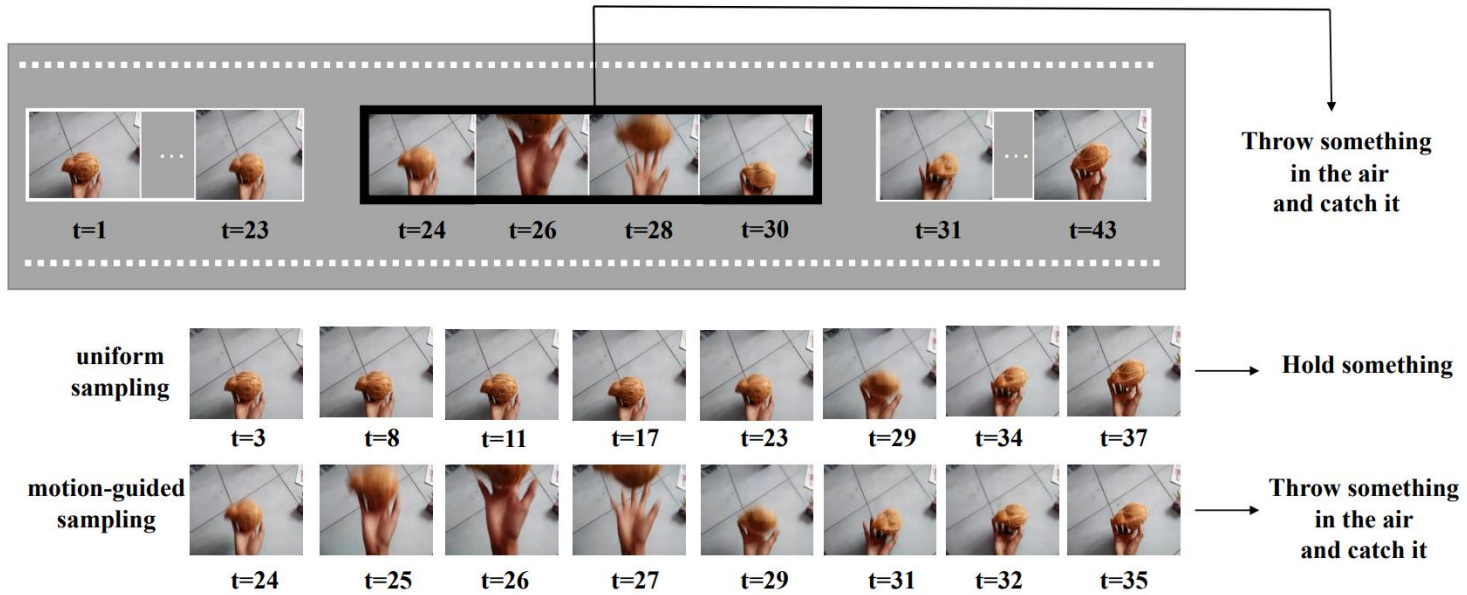
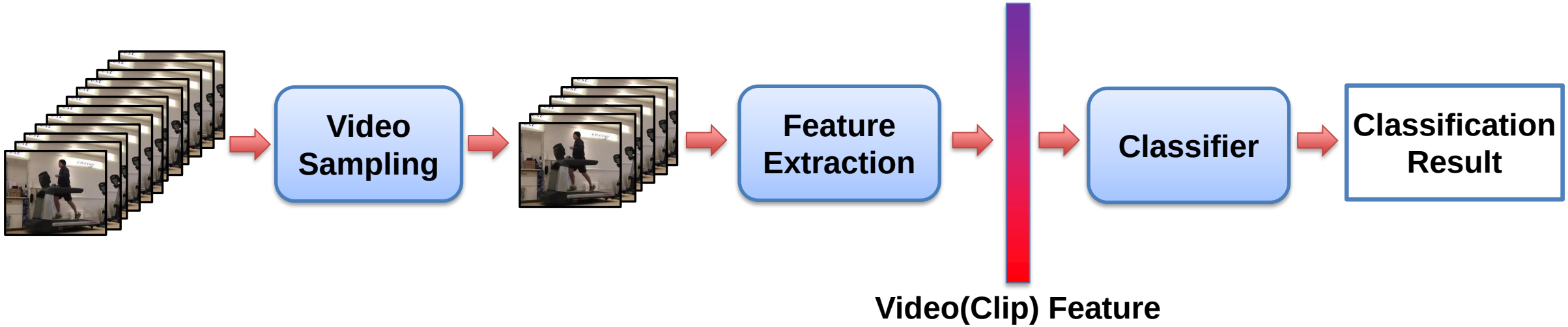


Figure 1. Sample eight frames from a video of throwing something in the air and catching it. Due to the quick moment in action, uniform sampling may miss the key information while our sampling strategy can identify and select frames with large motion magnitude.

Figure from paper

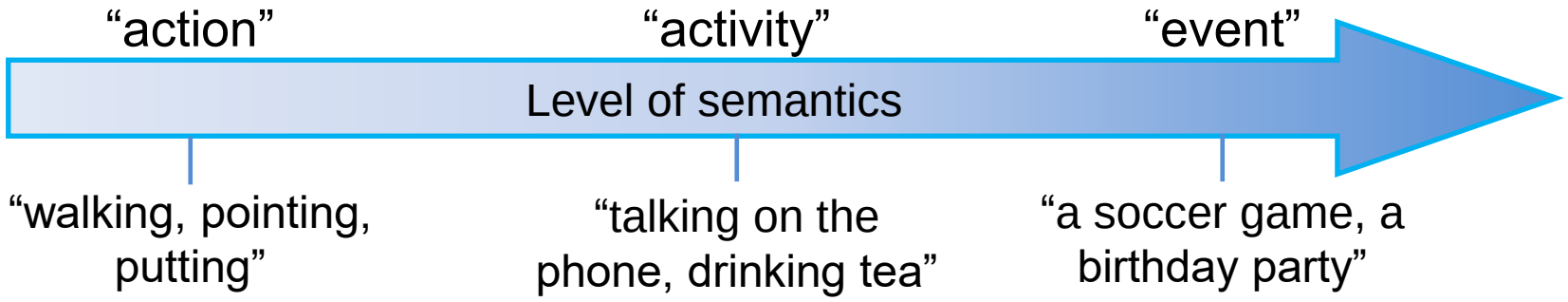
Video Fundamentals: Outline of Video Analysis



- Sample from the raw video input to form a “clip”.
- Apply feature extraction methods on the sampled “clip”, obtain the video (clip) feature.
  - Note: “Clip” and “Video” feature may be used interchangeably in literature. For simplicity we approximate the clip feature as the video feature.
- Apply a classifier (MAP decision rule/Bayes decision rule; Linear classifier; Nearest-neighbour classifier.
  - Classification is a general task category which includes (for videos): **human action recognition** (simplest), action segmentation, video semantic segmentation, etc.

Video Fundamentals: Human Action Recognition

- Given an input video, perform processing, and classify the human action in the video.



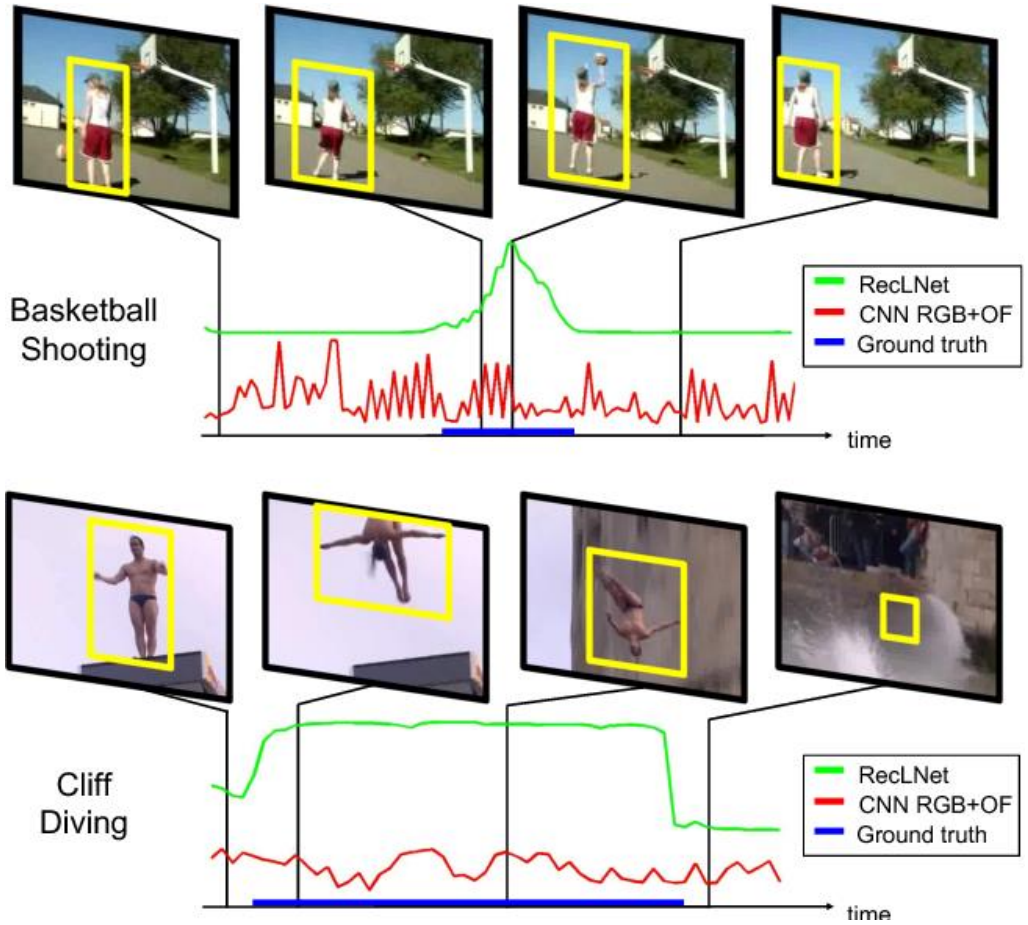
- Short clips, single action action, performed by human, video captured by ordinary camera.

```
Norm of logits: 138.468643

Top classes and probabilities
1.0 41.8137 playing cricket
1.49716e-09 21.494 hurling (sport)
3.84312e-10 20.1341 catching or throwing baseball
1.54923e-10 19.2256 catching or throwing softball
1.13602e-10 18.9154 hitting baseball
8.80112e-11 18.6601 playing tennis
2.44157e-11 17.3779 playing kickball
1.15319e-11 16.6278 playing squash or racquetball
6.13194e-12 15.9962 shooting goal (soccer)
4.39177e-12 15.6624 hammer throw
2.21341e-12 14.9772 golf putting
1.63072e-12 14.6717 throwing discus
```

Video Fundamentals: Other Video Tasks

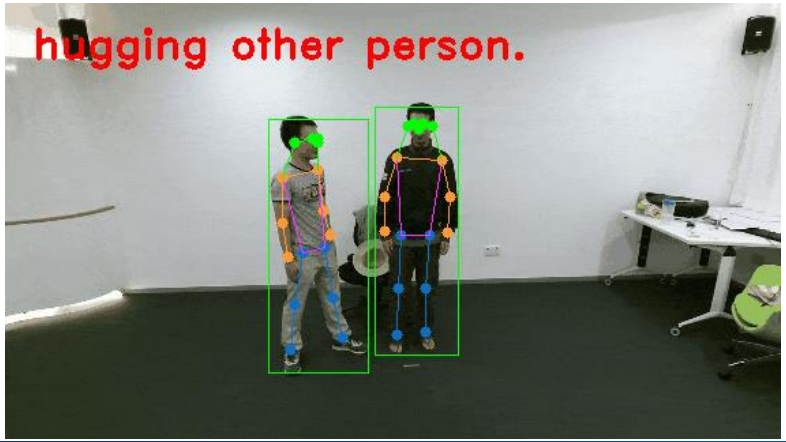
- Video Localization (Spatial/Temporal).



- Video Captioning.

Dense video captioning	A boy is riding a bicycle. He looses his balance and falls on the ground. He gets back up and starts riding again.
Single Sentence captioning	A boy in red t-shirt is riding a bicycle.

- Skeleton Action Recognition.



Video Fundamentals: Outline of Video Analysis

Raw video: Long, high FPS

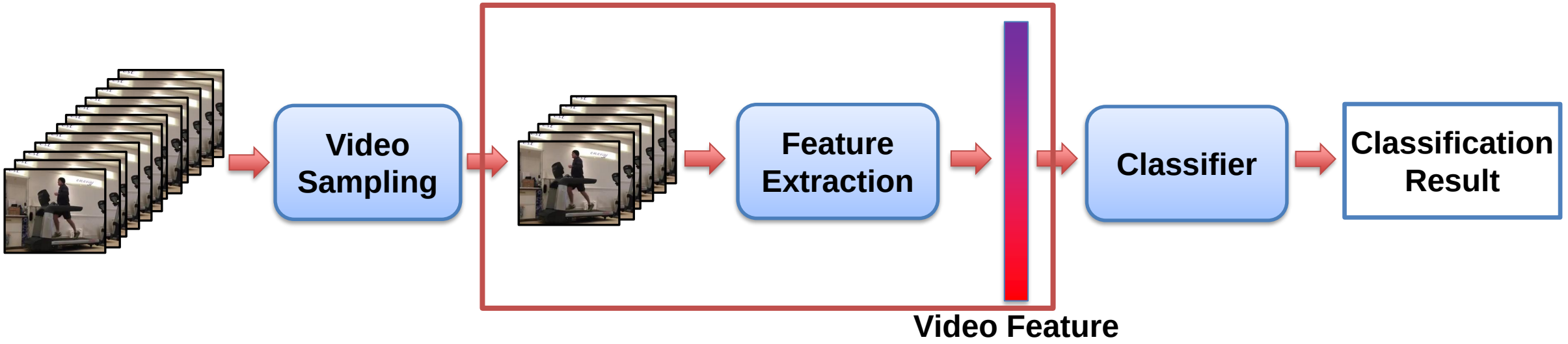
Training: Train model to classify short **clips** with low FPS

(Inference)

Testing: Run model on different clips, average predictions

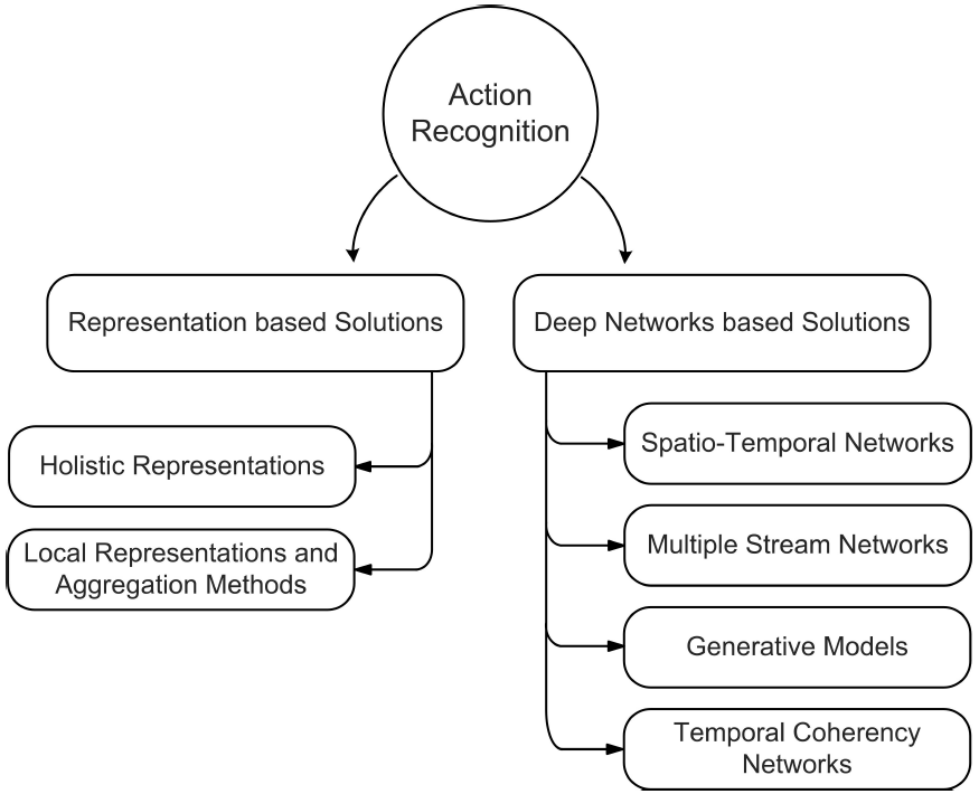
Figure from CS231N Stanford, Credit: Prof. Justin Johnson

Video Fundamentals: Outline of Video Analysis



- Effective feature extraction is the key towards accurate downstream tasks (e.g., classification).
- Feature extraction method categorized into **representation-based** and **learning-based**.
- Representation-based methods: extract video feature through handcrafted features.
- Learning-based methods: extract video feature by neural networks.
  - Often combine with downstream tasks and perform video analysis in an **end-to-end** manner.

Video Analysis 1: Representation-based Feature Extraction



- **Holistic** representations: Video feature extraction is performed by obtaining a **global** representation of human body structure, shape and movements.
  - Motion Energy Image (MEI) and Motion History Image (MHI).*
  - Optical Flow.*
- **Local** representations: Video feature is extracted locally, usually from local regions with salient motion information.
  - Space-Time Interest Points (STIP).
  - Motion Trajectory.

Figure credit: Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. Image and vision computing, 60, 4-21.

Video Analysis 1a: MEI

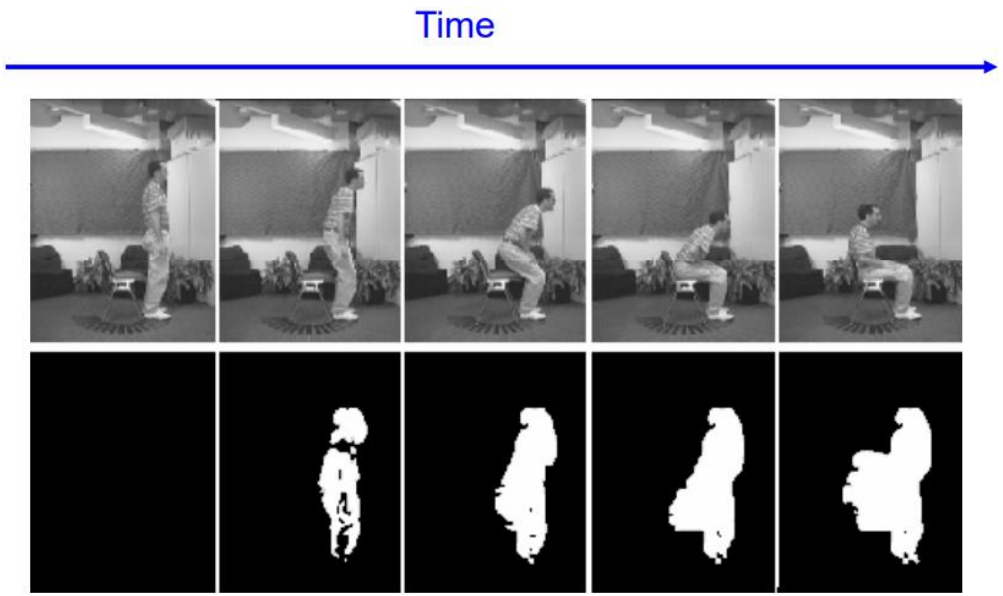


Figure credit: Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on pattern analysis and machine intelligence, 23(3), 257-267.

- MEI and MHIL: encode the motion-related information by a single image.
- MEI: spatial accumulation of motion.
- Mostly computed per-channel or in gray-scale
- MEI: represents “where” the motion occurs.
- Denote a frame at time t as $I(x, y, t)$, we obtain the frame difference between frame at time t and its previous adjacent frame as:

$$D(x, y, t) = \mathbb{1}_{Th}(I(x, y, t) - I(x, y, t-1))$$

$\mathbb{1}_{Th}$ is the indicator function formulated as:

$$\mathbb{1}_{Th}(x) = \begin{cases} 1 \; if \; x > Th \\ 0 \; if \; x < Th \end{cases}$$

- $D(x, y, t)$ is the binary image indicating regions of motion at time t.

Video Analysis 1a: MEI and MHI

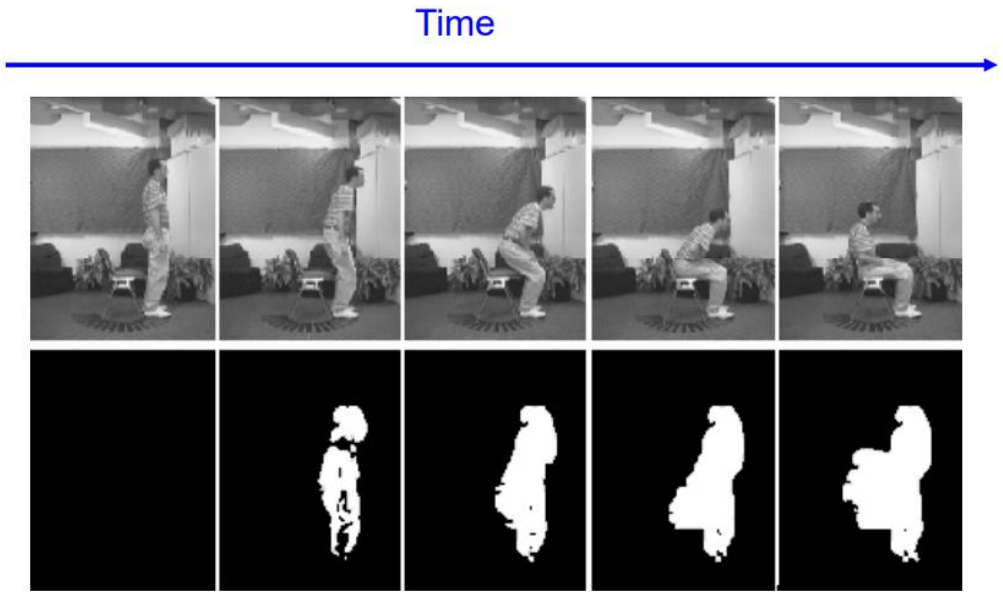


Figure credit: Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on pattern analysis and machine intelligence, 23(3), 257-267.

- The **binary** MEI corresponding to $I(x, y, t)$ is defined as:

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i)$$

- τ: Duration that defines the temporal extent of the motion.
- Does not encode "how", i.e., in what sequence, the motion happens.
- MHI: encode "how" the motion occurs. (**Not** binary image)
- In MHI, the pixel intensity is a function of the temporal history of motion at that point. For the results presented here, MHI use a simple replacement and decay operator:

$$H_\tau(x, y, t) = \begin{cases} \tau & if\ D(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t-1) - \delta) & otherwise \end{cases}$$

- $D(x, y, t) = 1$ indicates the motion is still ongoing.
- δ is the decay parameter (usually default to 1)

Video Analysis 1a: MEI and MHI

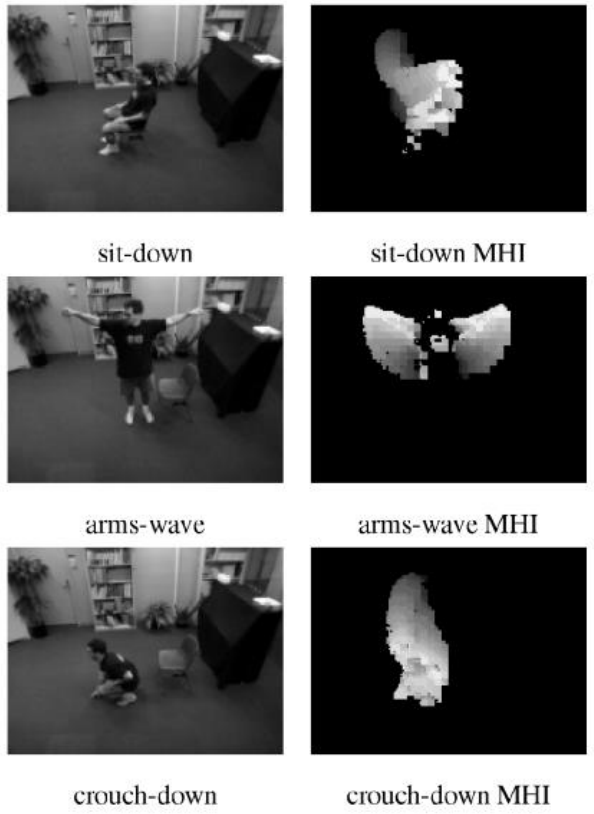


Fig. 4. Simple movements along with their MHIs used in a real-time system.

Figure credit: Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on pattern analysis and machine intelligence, 23(3), 257-267.

- MEI can be generated by thresholding MHI above zero.
- MEI + MHI → Temporal template.
- Disadvantage:
  - Limit to static background.
  - Limit to simple and short motion. (Videos in datasets at this point tend to only last tens of seconds)
  - Cannot cope with fine-grained actions.
  - Not view-invariant (motion overwriting problem).
  - Sensitive to parameter selection such as the variance of motion duration and decay parameter.

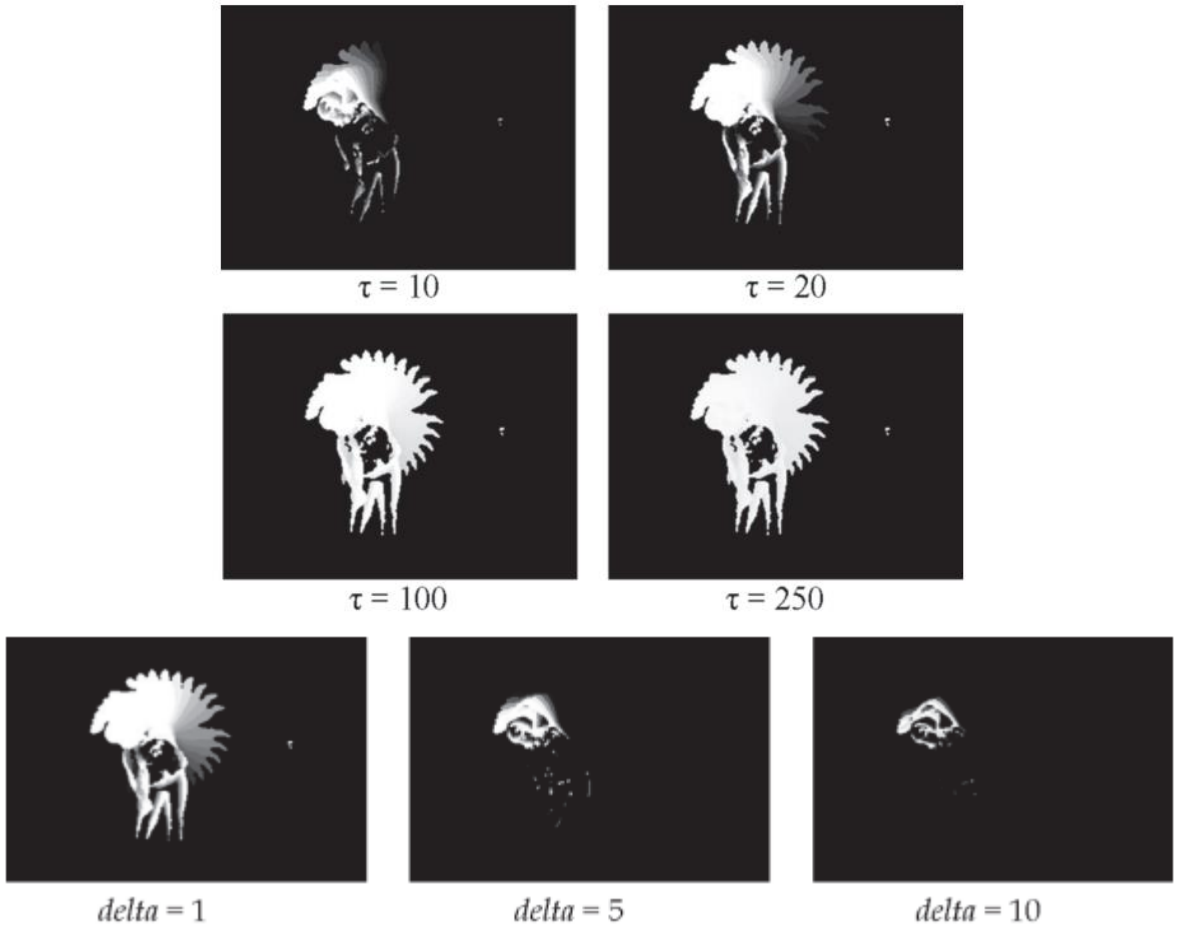
Video Analysis 1a: MEI and MHI

Figure 4: **Top:** A jumping sequence. **Middle:** The MEI template Bobick and Davis (2001). **Bottom:** The MHI template Bobick and Davis (2001). The MEI captures where the motion happens while the MHI template shows how the motion image is moving. The templates at the end of the action, shown in the rightmost column are used for representations.

Fig. 6 Examples of an input video frame, the corresponding motion energy image and motion history image computed by [15].

Figure credit: Left: Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. Image and vision computing, 60, 4-21. Right: Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. International Journal of Computer Vision, 130(5), 1366-1401.

Video Analysis 1a: MEI and MHI



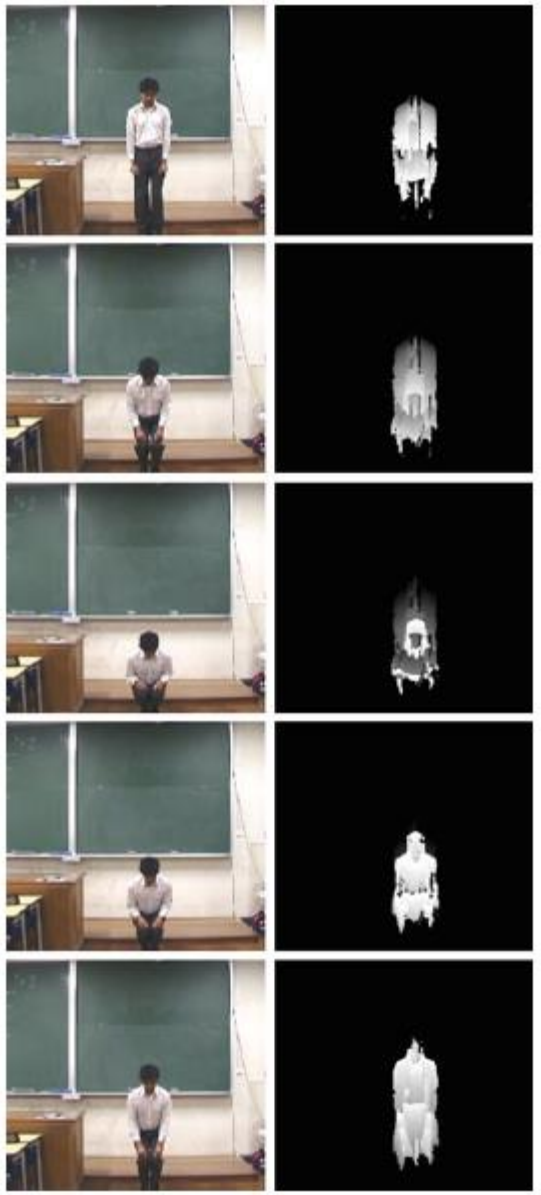


Fig. 10 Motion overwriting problem (due to self-occlusion) of the MHI method

Figure credit: (Left) Ahad, M. A. R. (2011). Computer vision and action recognition: A guide for image processing and computer vision community for action understanding (Vol. 5). Springer Science & Business Media.
(Right) Ahad, M., Rahman, A., Tan, J. K., Kim, H., & Ishikawa, S. (2012). Motion history image: its variants and applications. Machine Vision and Applications, 23(2), 255-281.

Video Analysis 1a: Extensions from MEI and MHI

Fig. 1. Space-time shapes of "jumping-jack," "walk," and "run" actions.

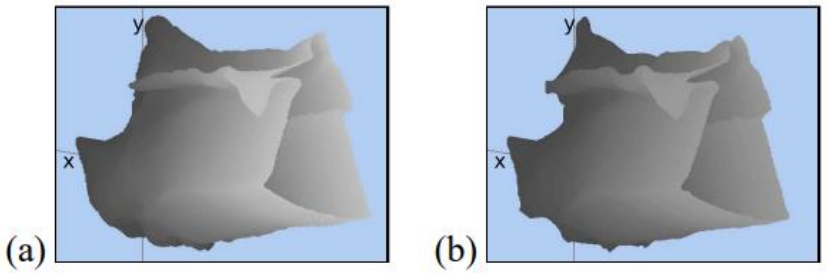
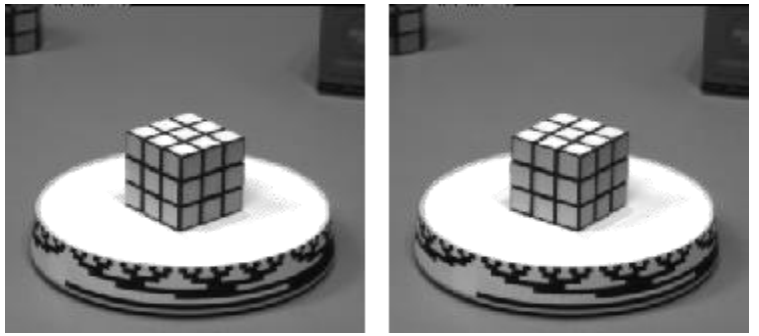


Figure 4: STVs for (a) dancer sequence with 40 frames, (b) synthetic dancer sequence with 20 frames, generated by randomly removing frames.

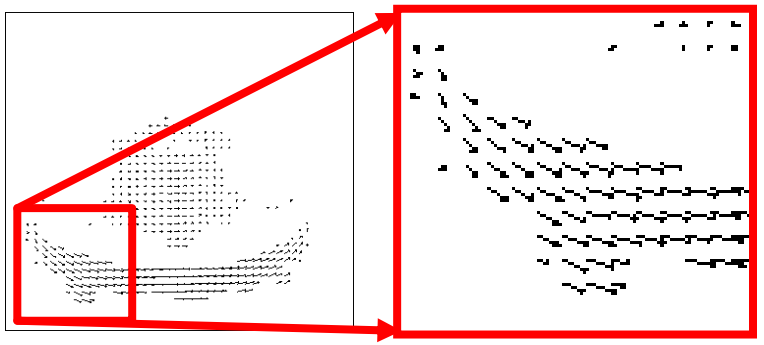
- Top: a volumetric extension of MEI.
  - Represent an action by a 3D shape induced from its outlines in the space-time.
  - The 3D shape is subsequently converted to 2D for analysis and downstream tasks.
- Bottom: Space-Time Volume (STV). (Can be viewed as a volumetric extension of MHI)
  - Build by stacking the object contours along the time axis.
  - Analyze STV by using the differential geometric surface properties, such as peaks, pits, valleys and ridges.

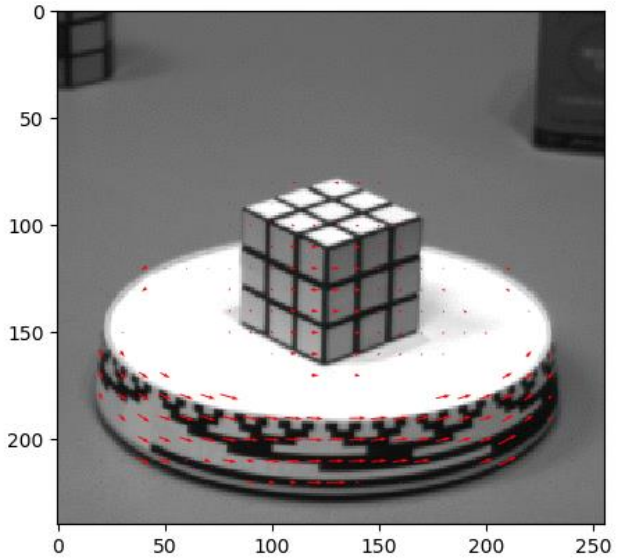
Figure credit: (Top) Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005, October). Actions as space-time shapes. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 2, pp. 1395-1402). IEEE.
(Bottom) Ahad, M., Rahman, A., Tan, J. K., Kim, H., & Ishikawa, S. (2012). Motion history image: its variants and applications. Machine Vision and Applications, 23(2), 255-281.

Video Analysis 1b: Optical Flow



- Instead of computing outlines or shape for action representation, **holistic** feature representation can also be extracted from **motion information** that are computed from videos.
- Optical flow is defined as the **apparent motion** of individual pixels on the image plane caused by the movement of object or camera. It is 2D vector field where each vector is a displacement vector showing the movement of points across consecutive frames.
- Apparent motion is the **appearance** of real motion from a sequence of still images.
- Ideally, optical flow would be the same as the actual motion field.

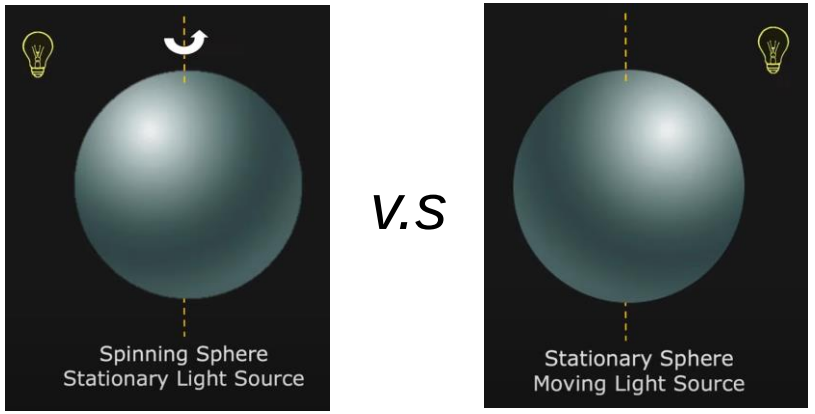


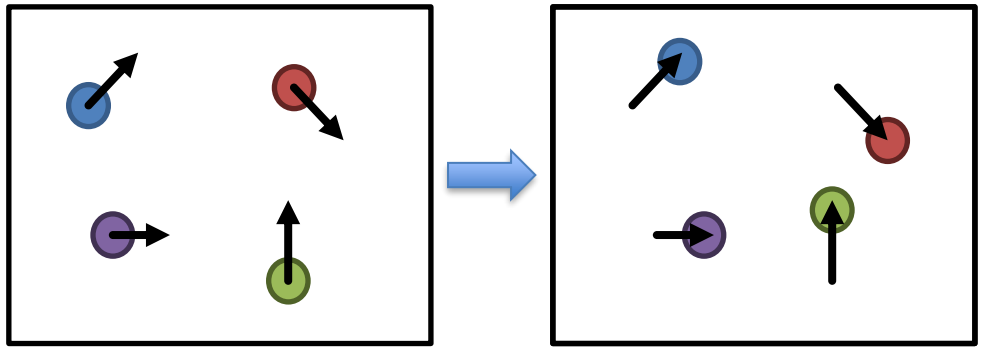


Top figure from EECS 442 University of Michigan, Credit: Prof. Justin Johnson

Video Analysis 1b: Optical Flow

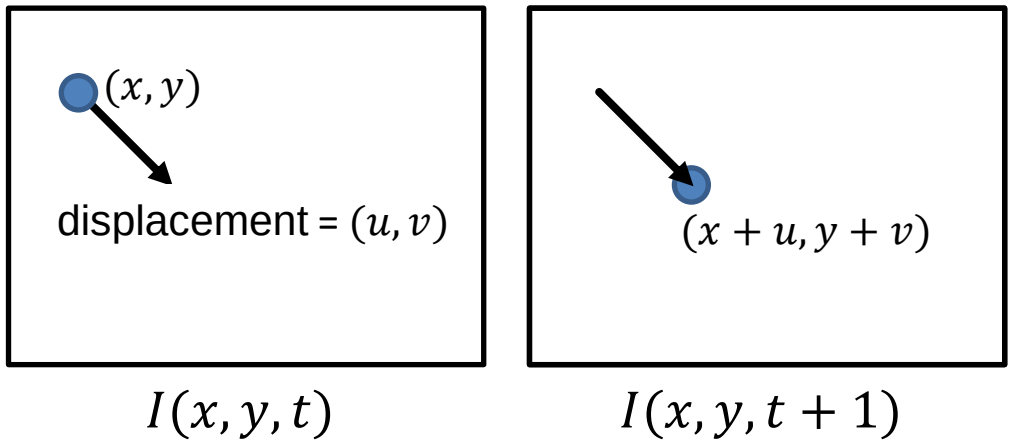
- **Note: apparent motion can be caused by lighting changes without any actual motion. (think of shadows)**
- A rotating sphere with a static light source produces a static image, but a stationary sphere with a moving light source produces drifting intensities.
- Start by estimating motion of each pixel separately, then consider motion of entire image.
- Estimating optical flow is formulated as:
- Given two subsequent frames, estimate the apparent motion field $u(x, y)$ and $v(x, y)$ between them.
- At the pixel level, we need to solve the following correspondence problem.
- Given pixel at time t, find nearby pixels of the same color at time $t + 1$





Video Analysis 1b: Optical Flow

- Key assumptions
  - **Brightness constancy**: projection of the same point looks the same in every frame.
  - **Small motion**: points do not move very far.
  - **Spatial coherence**: points move like their neighbors.



- Brightness constancy:

$$I(x, y, t) = I(x + u, y + v, t + 1) \quad \text{Eq. B}$$

- It is impossible to find corresponding pixels in a brute force manner.
- Linearize the right using Taylor expansion with the first-order Taylor series approximation (I_x denotes gradient of I over x)

$$I(x, y, t) \approx I(x, y, t + 1) + I_x u + I_y v$$

$$\boxed{I(x, y, t + 1) - I(x, y, t)} + I_x u + I_y v = 0$$

Essentially I_t

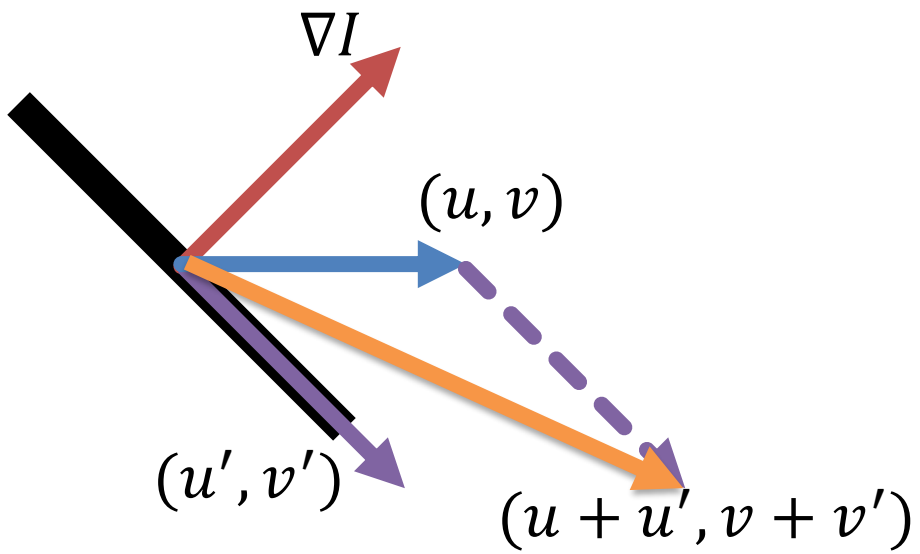
$$I_x u + I_y v + I_t = 0$$

or equivalently $\nabla I \cdot (u, v) + I_t = 0$

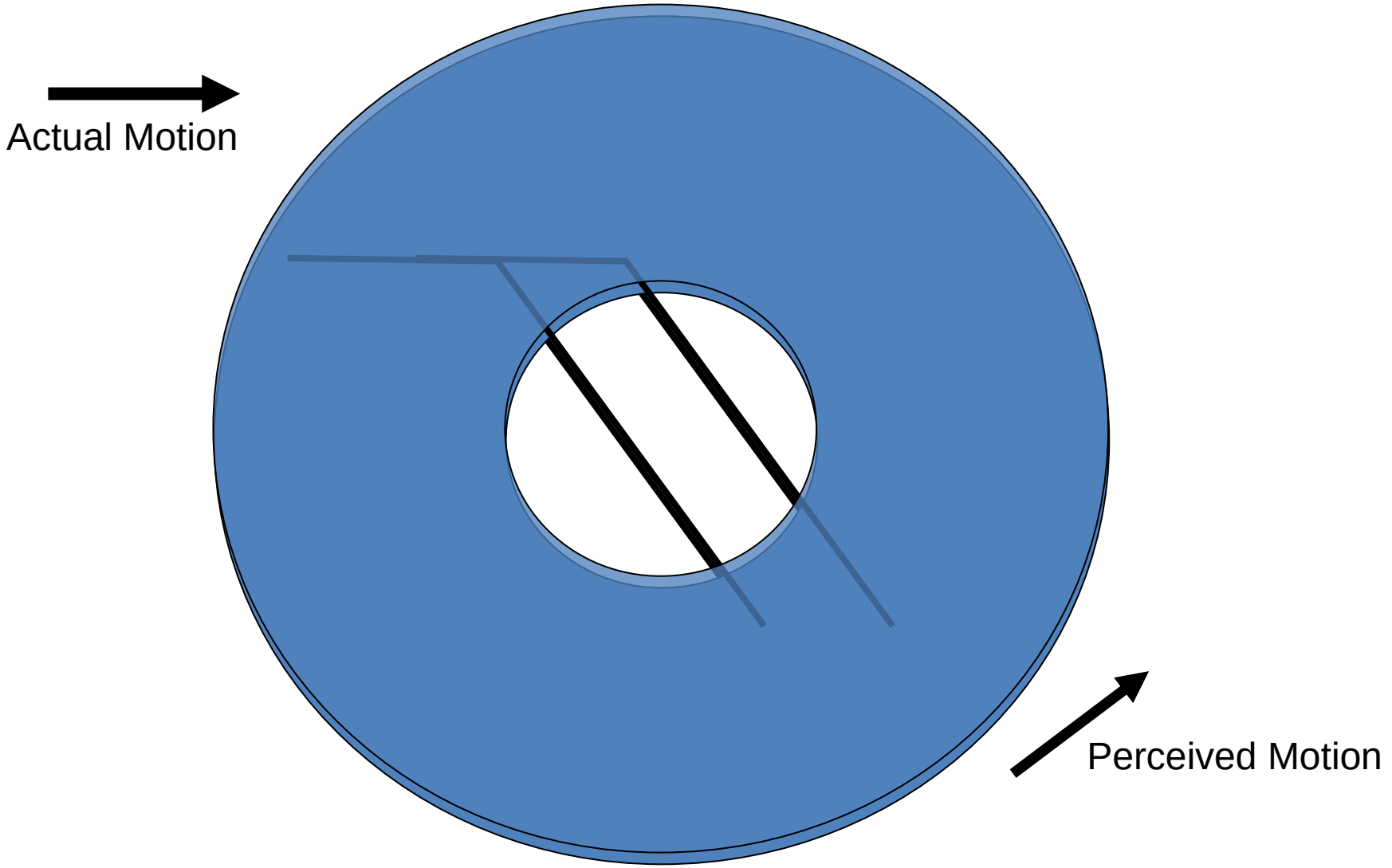
Taylor expansion: $f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots$

Video Analysis 1b: Optical Flow – Brightness Constancy

- What does the equation $I_x u + I_y v + I_t = \nabla I \cdot (u, v) + I_t = 0$ indicates?
- A single equation but two unknowns – there are trivial solutions $\nabla I \cdot (u, v) = 0$.
- Brightness constancy constraint can only identify the motion along the gradient but **not** the motion perpendicular to the gradient.
- If (u, v) satisfies brightness constancy constraint, so does $(u + u', v + v')$ if $\nabla I \cdot (u', v') = 0$.



Video Analysis 1b: Optical Flow – Aperture Problem



Video Analysis 1b: Optical Flow

- How to solve the aperture problem?
  - Prevent trivial solutions by getting more equations for a set of unknowns.
- Spatial coherence constraint: assume the pixel's neighbors have the same (u, v) – **Lucas-Kanade** [1] optical flow.
  - E.g., 5×5 window gives 25 new equations.

$$I_t + I_x u + I_y v = 0$$

$$\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ \vdots & \vdots \\ I_x(p_{25}) & I_y(p_{25}) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(p_1) \\ \vdots \\ I_t(p_{25}) \end{bmatrix}$$

$$p_1 = (x_1, y_1)$$

[1] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 674–679, 1981.

Video Analysis 1b: Lucas-Kanade Optical Flow

- $\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ \vdots & \vdots \\ I_x(p_{25}) & I_y(p_{25}) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(p_1) \\ \vdots \\ I_t(p_{25}) \end{bmatrix}$ is essentially in the form of: $\underset{25\times 2}{\boldsymbol{A}} \quad \underset{2\times 1}{\boldsymbol{d}} = \underset{25\times 1}{\boldsymbol{b}}$
- The solution is given by $(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{d} = \boldsymbol{A}^T\boldsymbol{b} \rightarrow \boldsymbol{d} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$.
- In matrix form, we have:
- $\begin{bmatrix} \sum \boldsymbol{I_x I_x} & \sum \boldsymbol{I_x I_y} \\ \sum \boldsymbol{I_x I_y} & \sum \boldsymbol{I_y I_y} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum \boldsymbol{I_x I_t} \\ \sum \boldsymbol{I_y I_t} \end{bmatrix}$
- $\boldsymbol{A}^T\boldsymbol{A}$ $\boldsymbol{A}^T\boldsymbol{b}$

(summation over all pixels in the window)

- The next question becomes when can we find the solution (u, v)

Video Analysis 1b: Lucas-Kanade Optical Flow

- Define $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} \sum \boldsymbol{I_x I_x} & \sum \boldsymbol{I_x I_y} \\ \sum \boldsymbol{I_x I_y} & \sum \boldsymbol{I_y I_y} \end{bmatrix}$, which is a second moment matrix, and can be decomposed (through eigenvector decomposition) as $\boldsymbol{M} = \boldsymbol{R}^{-1} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \boldsymbol{R}$.
- Estimation of optical flow is well-conditioned for regions with high "cornerness", conditioned by a) $\boldsymbol{\lambda_1}$ **and** $\boldsymbol{\lambda_2}$ **are large**, b) $\boldsymbol{\lambda_1 \sim \lambda_2}$, and c) M is invertible (therefore cannot be precisely equal brightness → Hight texture region.
- λ_1 and λ_2 are eigenvectors.

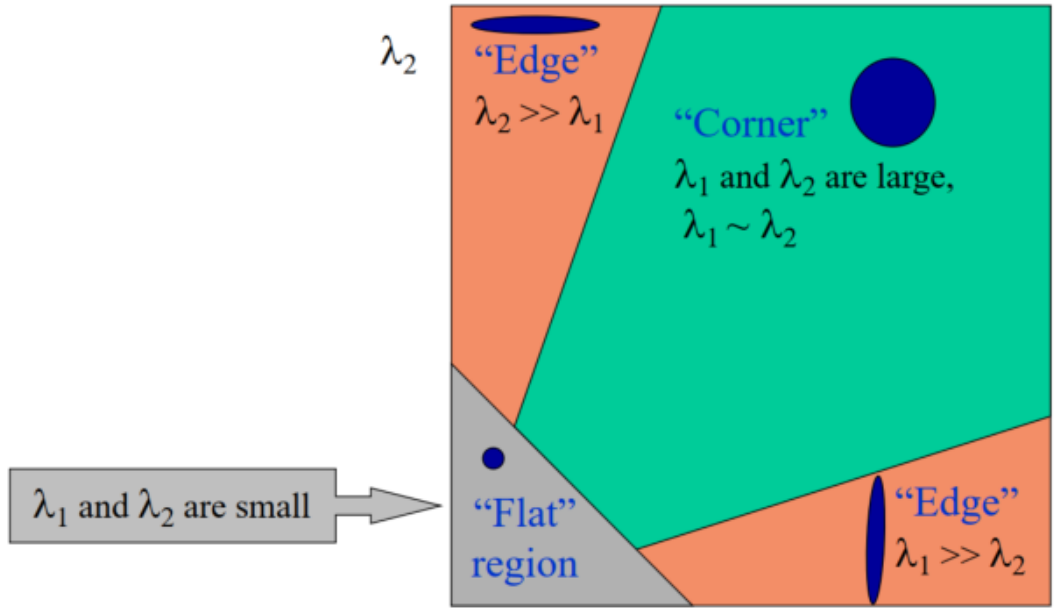
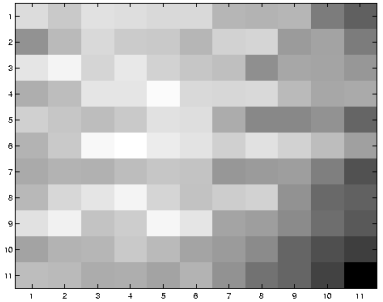
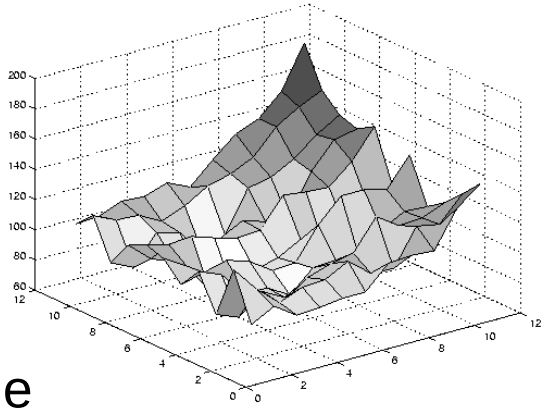


Figure from CS 543/ECE 549 University of Illinois Urbana-Champaign, Credit: Prof. Saurabh Gupta

Video Analysis 1b: Lucas-Kanade Optical Flow

- Failed case 1: Low-texture regions.





$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \nabla \mathrm{I} (\nabla \mathrm{I})^{\mathrm{T}}$$

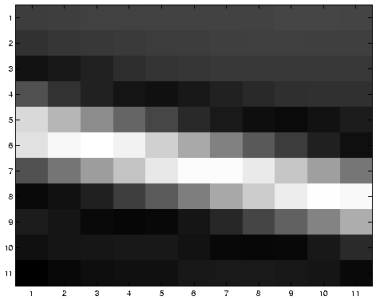
– Gradients have small magnitude

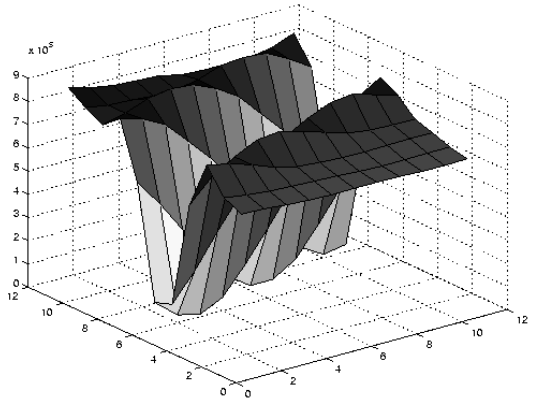
– Small λ_1, small λ_2

Figure from EECS 442 University of Michigan, Credit: Prof. Justin Johnson

Video Analysis 1b: Lucas-Kanade Optical Flow

- Failed case 2: Edges.





$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \nabla \mathrm{I} (\nabla \mathrm{I})^{\mathrm{T}}$$

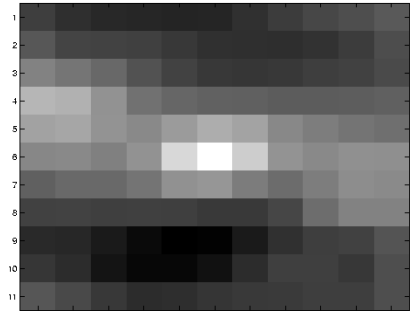
– Large gradients, along a certain direction

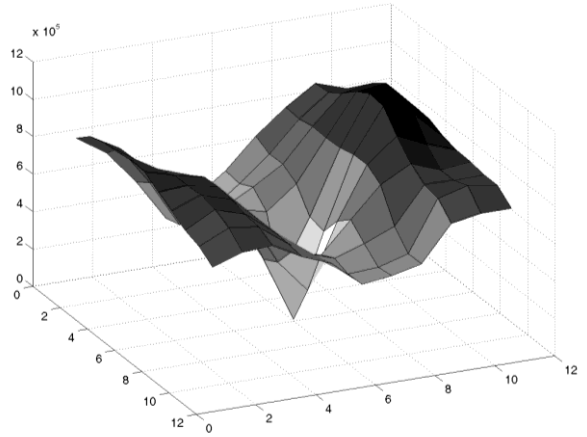
– Large λ_1, small λ_2, or vice versa

Figure from EECS 442 University of Michigan, Credit: Prof. Justin Johnson

Video Analysis 1b: Lucas-Kanade Optical Flow

- Success case: High-texture regions.





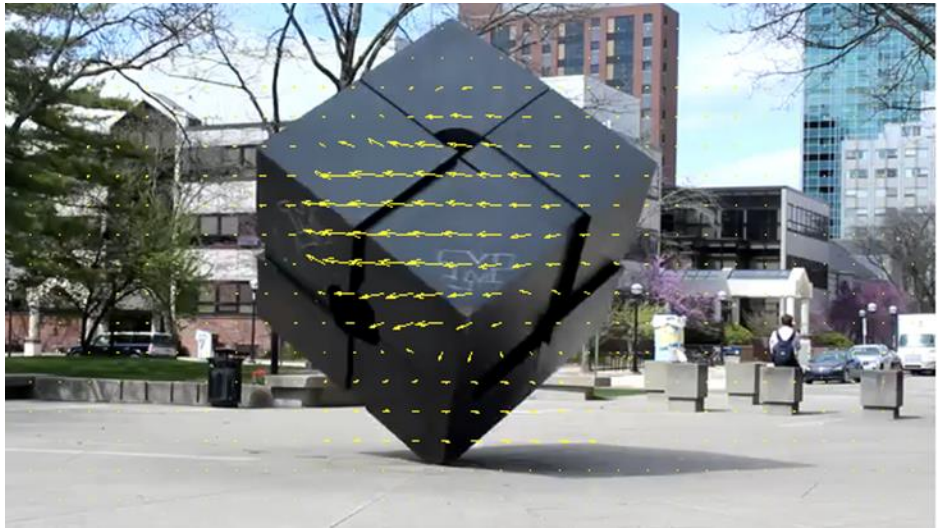
$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \nabla \mathrm{I} (\nabla \mathrm{I})^{\mathrm{T}}$$

– Gradients are different across different directions, large magnitudes

– Large λ_1, large λ_2

Figure from EECS 442 University of Michigan, Credit: Prof. Justin Johnson

Video Analysis 1b: Lucas-Kanade Optical Flow



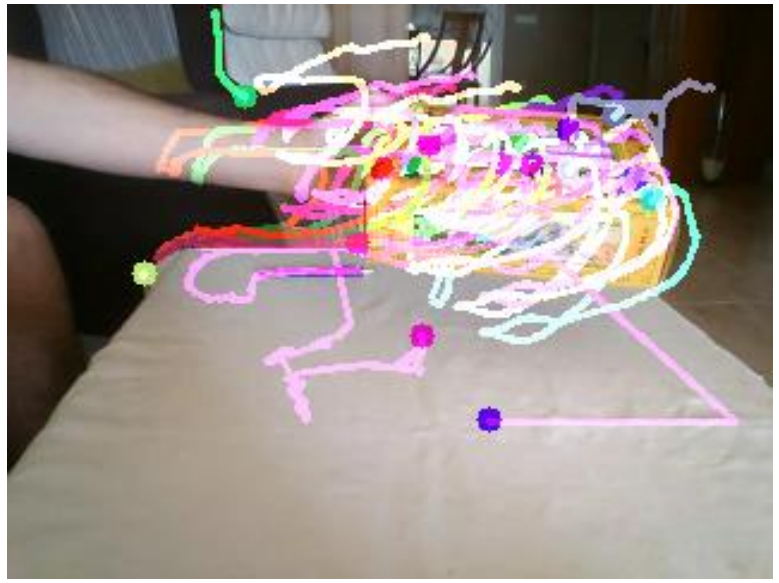


Figure from MATLAB and OpenCV tutorials

Video Analysis 1b: Lucas-Kanade Optical Flow

- When would Lucas-Kanade fail?
  - Its success is conditioned by a) brightness constancy; b) small motion; c) spatial coherence.
- Brightness constancy does not hold:
  - Other form of pixel and feature matching (e.g., SIFT).
- The motion is large (larger than a certain number of pixels):
  - Multi-resolution estimation, iterative refinement.
- Point doesn’t move like neighbors:
  - Figure out which points move together, then come back and fix.

Video Analysis 1b: Multi-Resolution Estimation

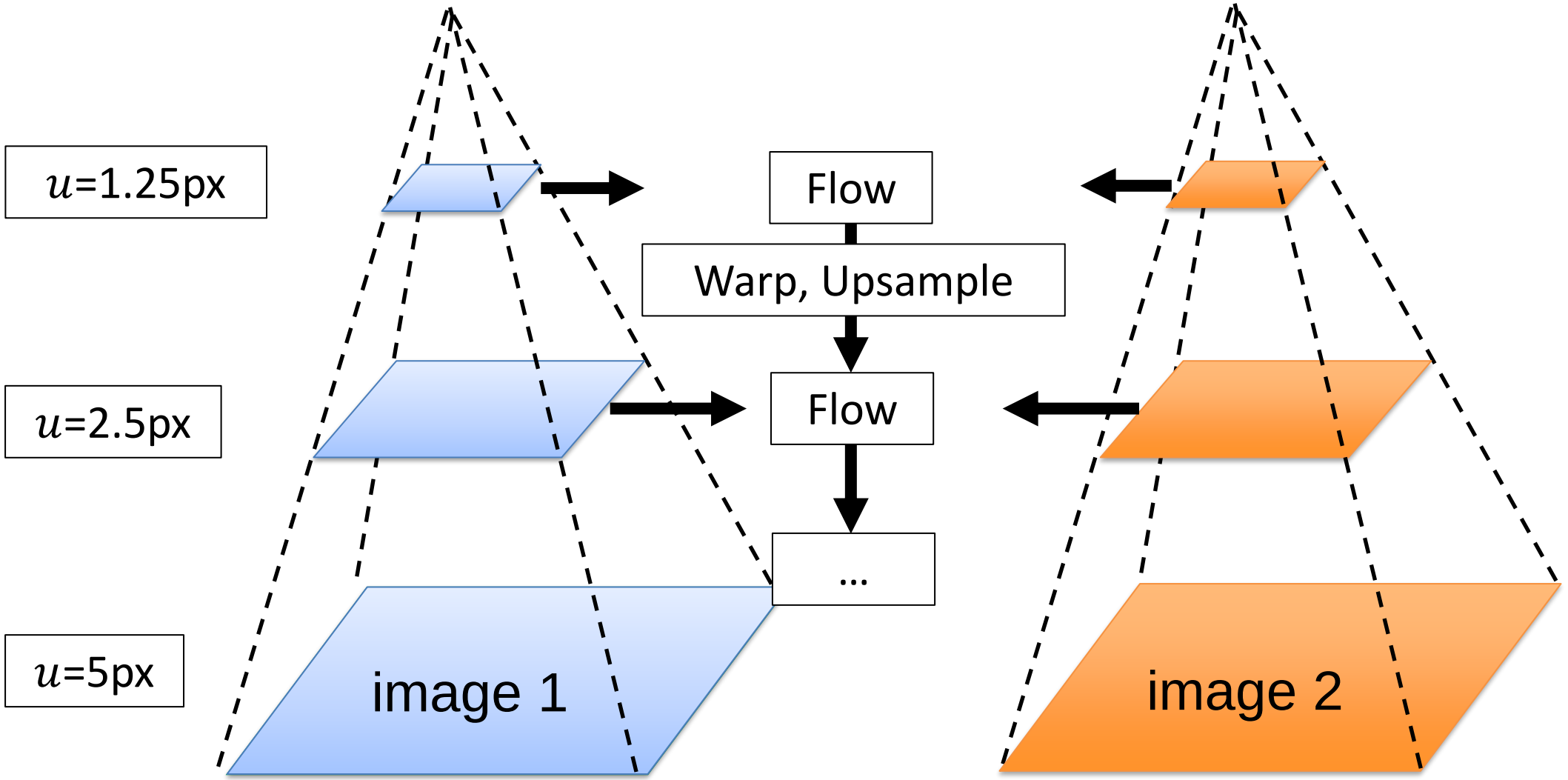
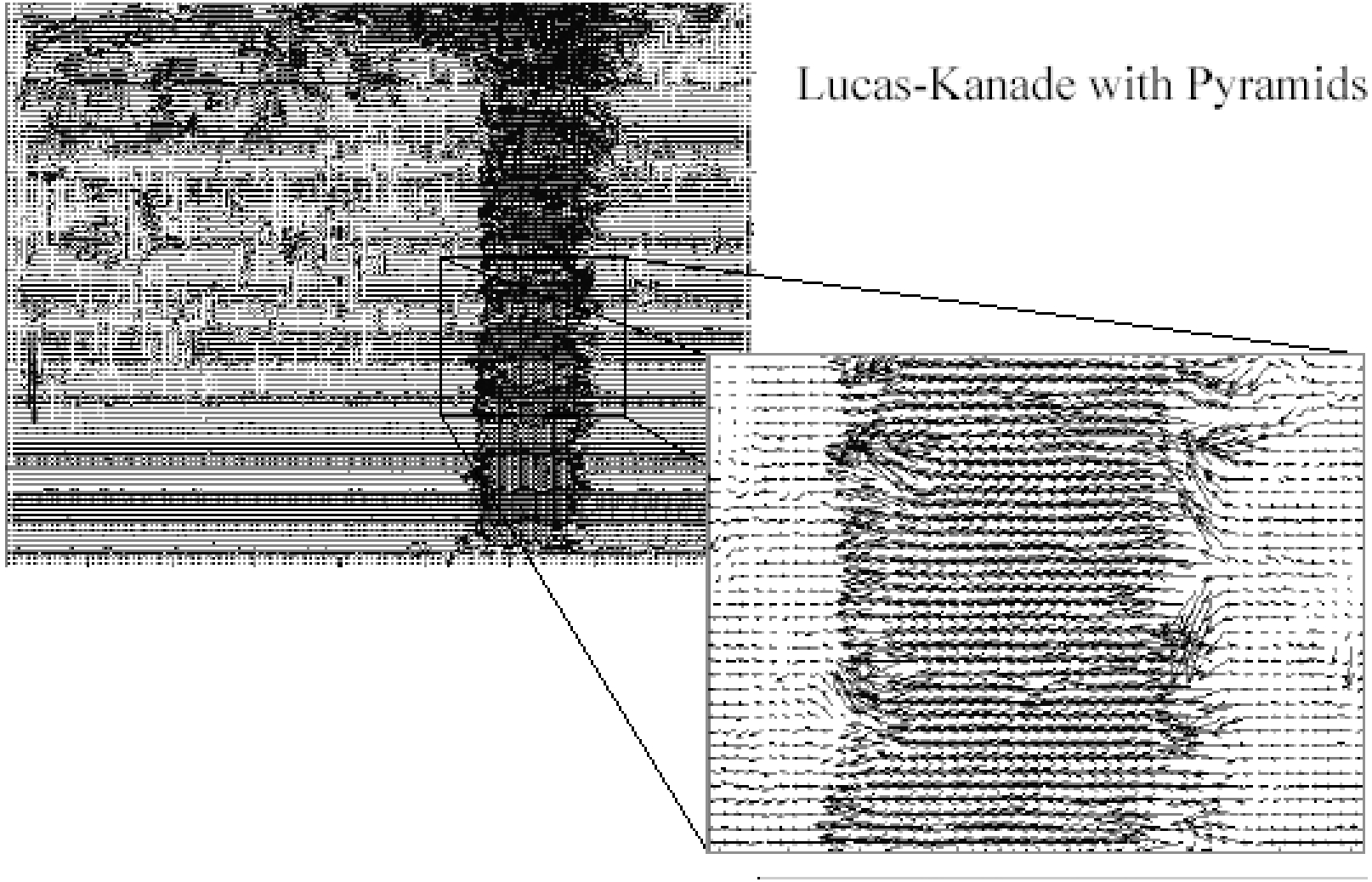


Figure from EECS 442 University of Michigan, Credit: Prof. Justin Johnson

Video Analysis 1b: Multi-Resolution Estimation



Video Analysis 1b: Motion Segmentation

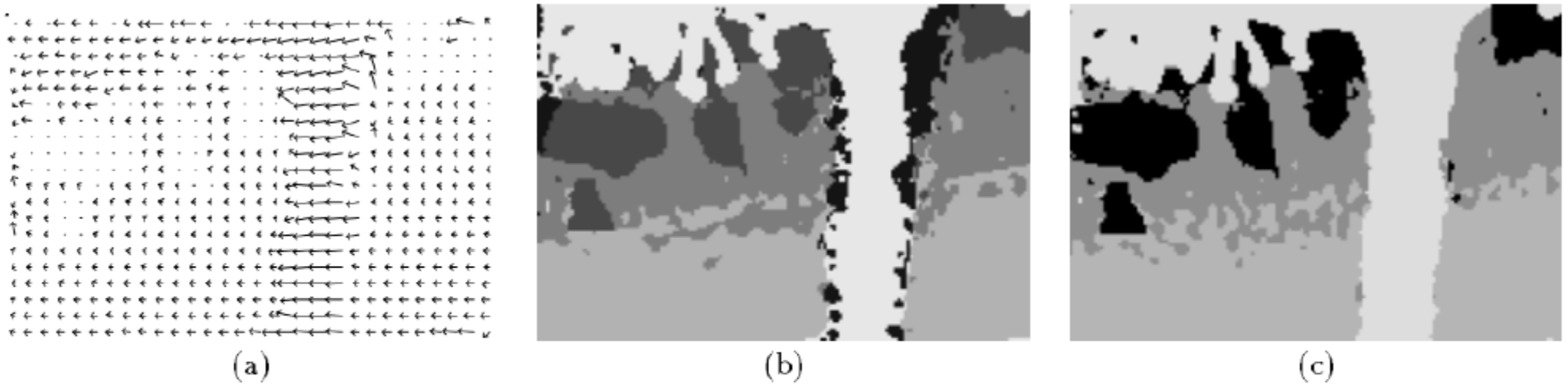


Figure 11: (a) The optic flow from multi-scale gradient method. (b) Segmentation obtained by clustering optic flow into affine motion regions. (c) Segmentation from consistency checking by image warping. Representing moving images with layers.

Figure credit: Wang, J. Y., & Adelson, E. H. (1994). Representing moving images with layers. IEEE transactions on image processing, 3(5), 625-638.

Video Analysis 1b: Open Questions on Optical Flow

- Why is Optical Flow still leveraged even until today?
- What are the disadvantages of Optical Flow?

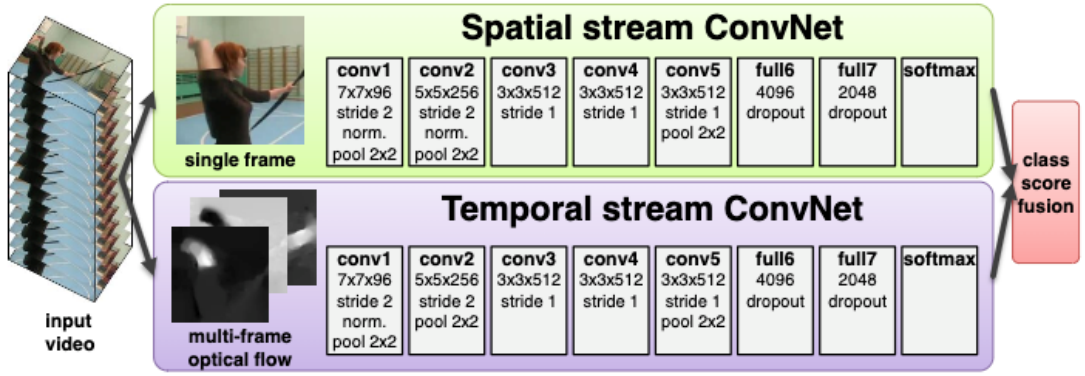
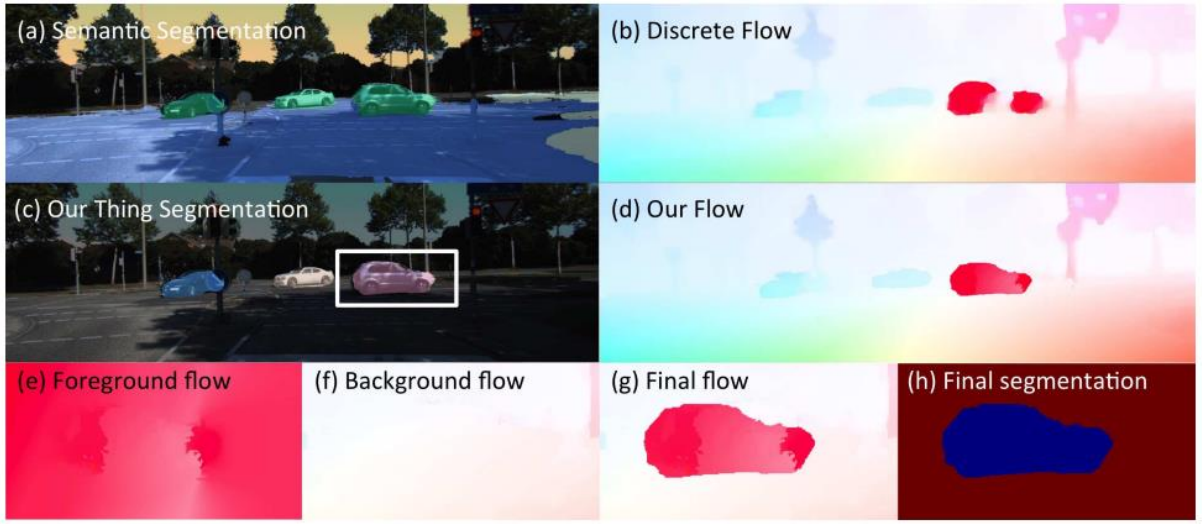


Figure 1: **Two-stream architecture for video classification.**



Video Analysis 1c: Local representations

- Video feature extracted locally, usually from local regions with salient motion information.
- Examples: Space-Time Interest Points (STIP); Motion Trajectory (Sparse/Dense Trajectory).
- STIP: The key points (interest points) are points with **high data variation in its space-time neighborhood** → motion information.
- Similar to “Corner Detection”
- STIP: Extend the Harris Corner Detection from 2D to 3D (with temporal dimension) – find local extremas with cuboids.

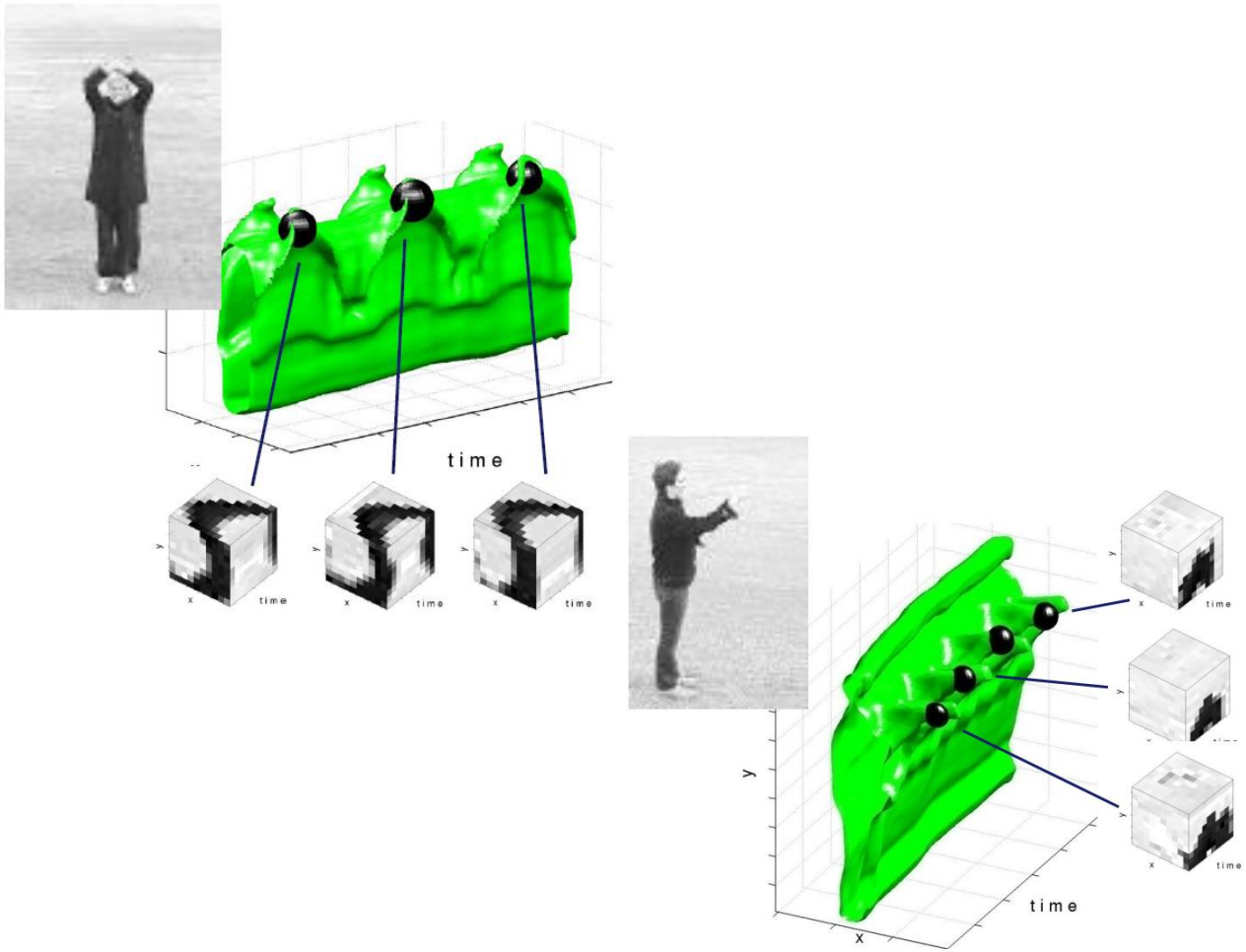


Figure from RECVIS 10 (2010) INRIA France, Credit: Dr. Ivan Laptev

Video Analysis 1c: STIP

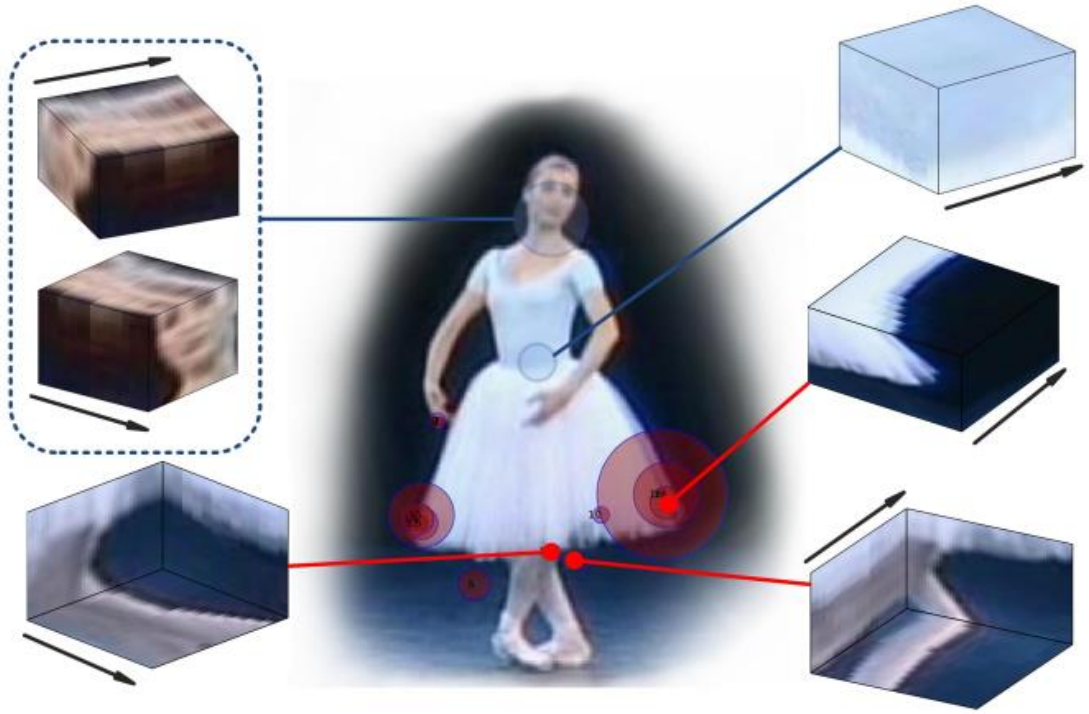
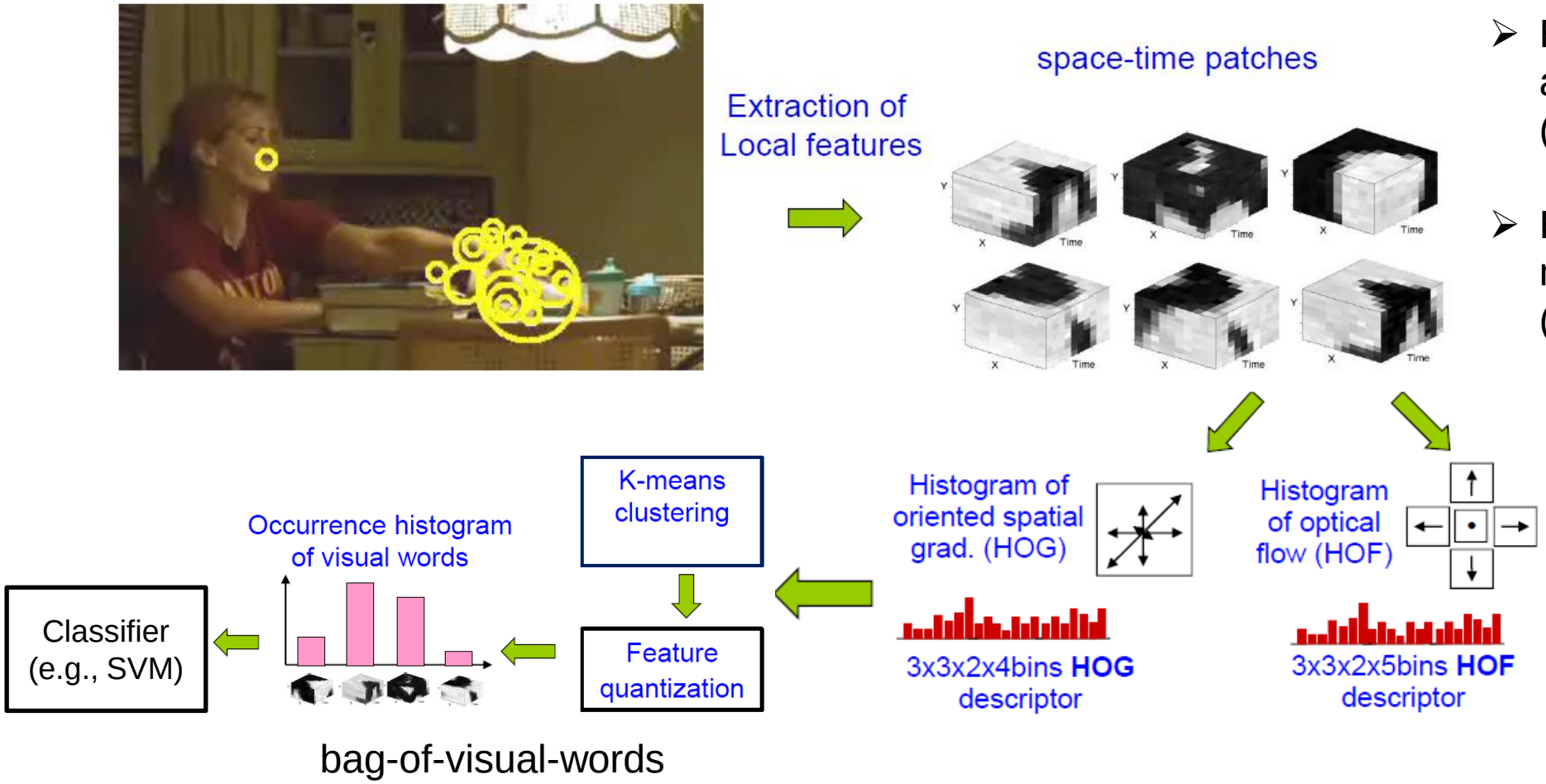


Figure 6: Marked in red are the detected spatiotemporal interest points of Laptev (2005). Spatial changes along the time axis (marked with an arrow) are noticeable. In this ballet video, the dancer keeps her head still throughout the video. Hence, despite having significant amount of spatial features, no spatiotemporal interest point is detected on the face. Similarly, in her waist no spatiotemporal interest point can be detected as a result of limited spatial variations.

Figure credit: Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. Image and vision computing, 60, 4-21.

Video Analysis 1c: STIP

- Process of recognizing actions with STIP:

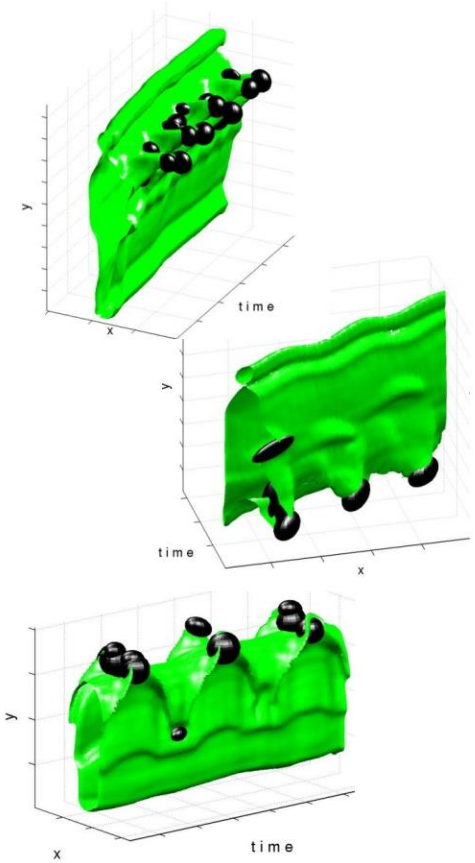


- HOG focuses on static appearance information. (gradient of appearance)
- HOF captures the local motion information. (gradient across time)

Figure from UIUC and Advanced Digital Sciences Center, Singapore

Video Analysis 1c: STIP – Bag-of-Visual-Words

- Bag-of-Visual-Words (BOVW): extending the concept of Bag-of-Words (BOW) to visual:
  - BOW: count the number of each word (from a dictionary) in a document, use the frequency of each word to construct the feature of the document.
  - BOVW: keypoints and their descriptors as “visual words”, the set of keypoints and descriptors are used to construct dictionaries and represent the video with the frequency of keypoints and their descriptors.



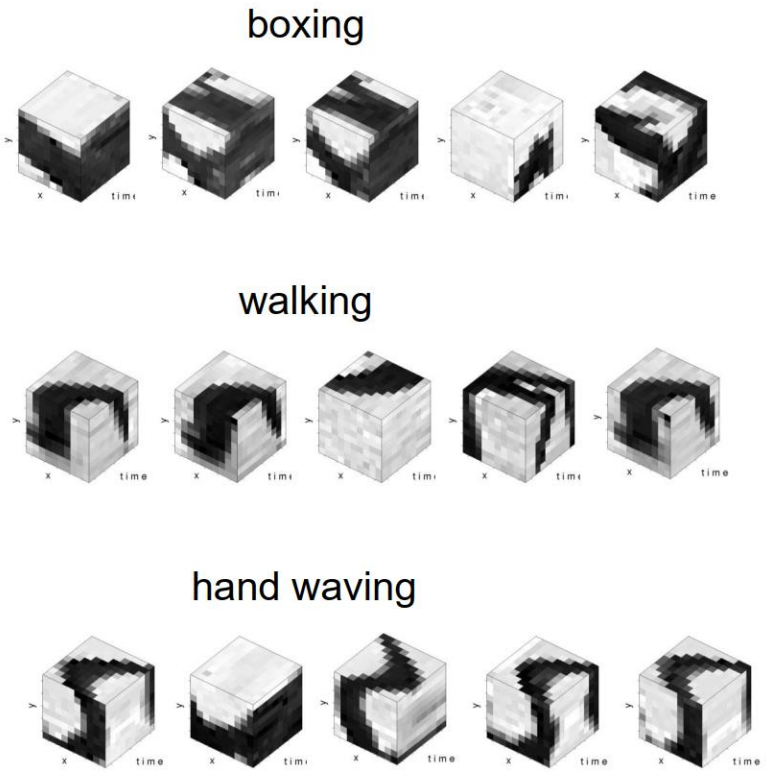


Figure from RECVIS 10 (2010) INRIA France, Credit: Dr. Ivan Laptev