Paper: **A robust and efficient video representation for action recognition**

Authors: Heng Wang,  Dan Oneata,  Jakob Verbeek,  Cordelia Schmid

# Problem addressed using proposed framework

- Action Recognition

(a) answer-phone (a) get-out-car (a) fight-person (b) push-up (b) cartwheel (b) sword-exercise

- Action Localization

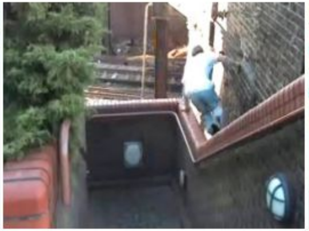
(g) drinking (g) smoking (h) sit-down (h) open-door

- Event recognition

(i) changing-vehicle-tire

(i) unstuck-vehicle

(i) making-a-sandwich


(i) parkour

(i) grooming-an-animal

(i) flash-mob-gathering

# Introduction

Diversity of realistic video data has lead to following challenges:

- Large magnitude of intra-class variation.By,factors such as style and duration.

- Challenges of images like background clutter, occlusions.Video challenges like motion clutter, variability in camera motion.

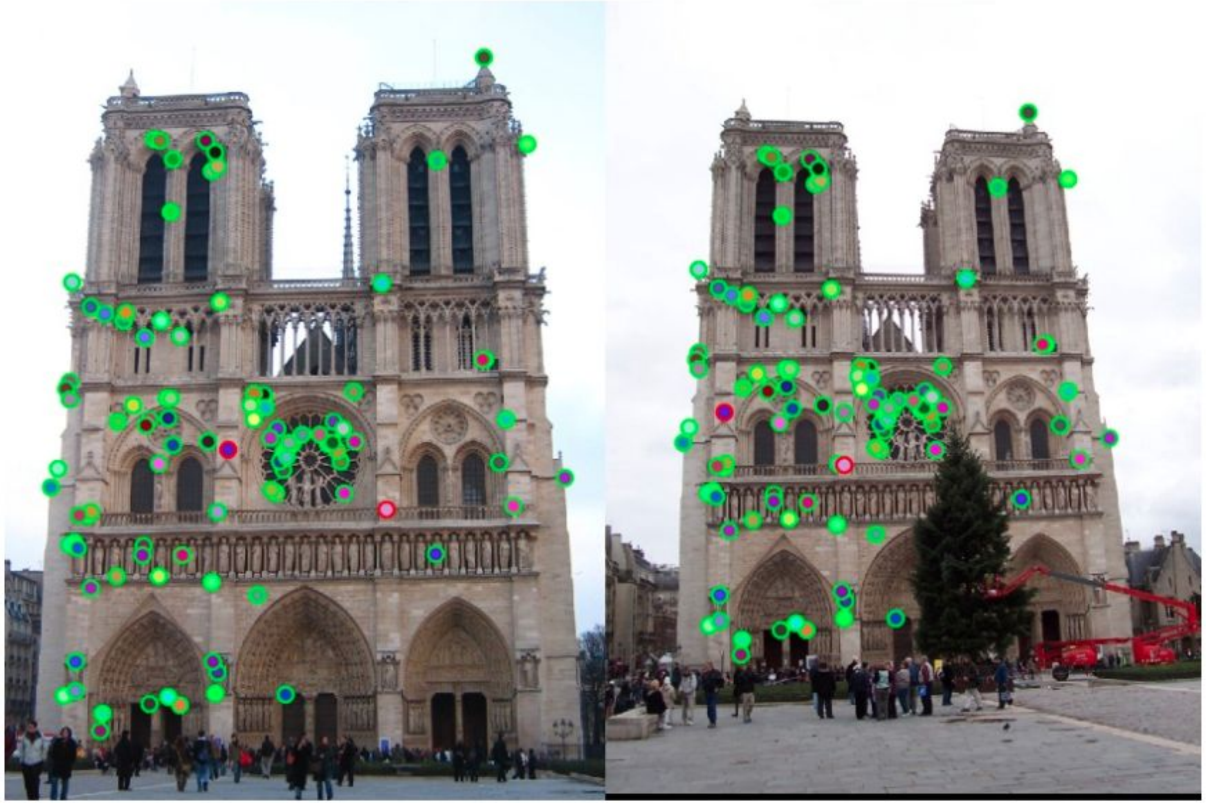
- Processing large amount of video data in the dataset.

# Aim of paper

- This paper introduces a state-of-the-art video representation and applies it to efficient action recognition and detection.

- To, improve the dense trajectory features camera motion estimation is needed.

Keypoints in an image

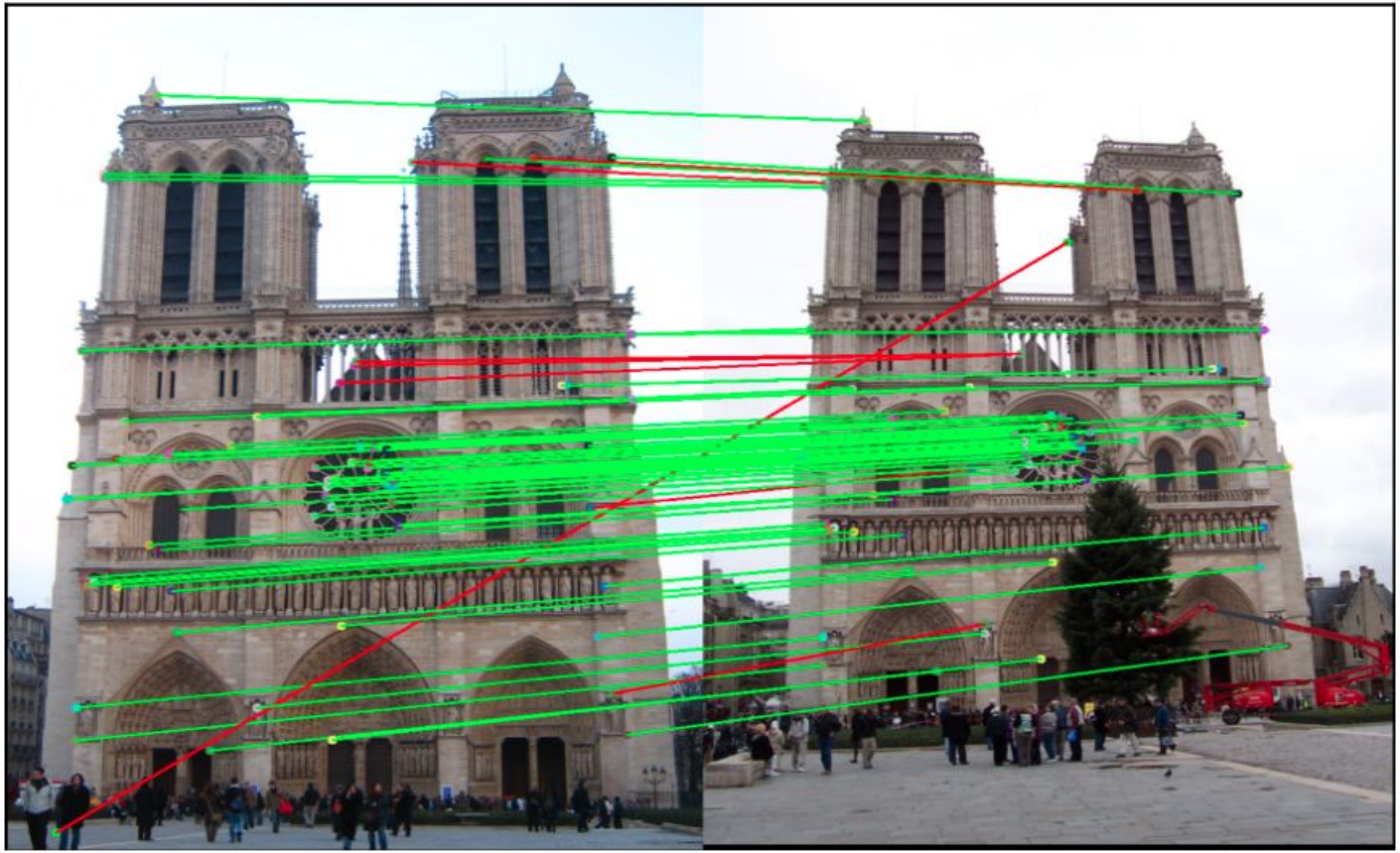


Properties of Interest point:

- Well-defined position in image space.
- *stable* under local and global perturbations i.e. computed with high degree of repeatability.
- Provide efficient detection

References:
https://medium.com/@deepanshut041/introduction-to-feature-detection-and-matching-65e27179885d

# Feature Matching example



References:
https://medium.com/@deepanshut041/introduction-to-feature-Detection-and-matching-65e27179885d

## Remove camera motion from the frames

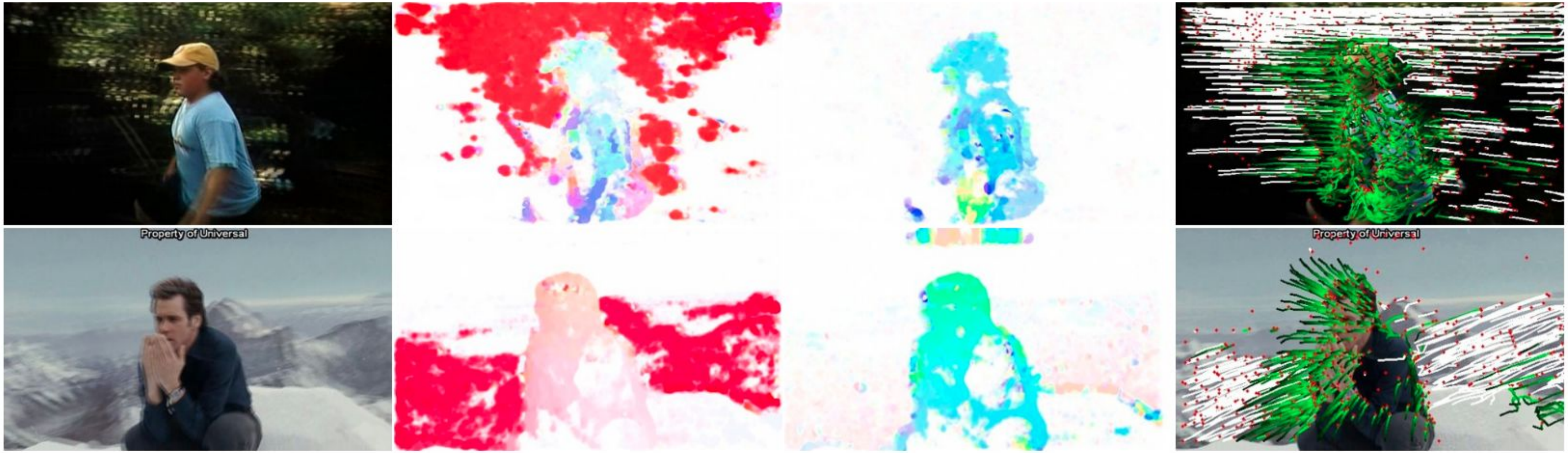
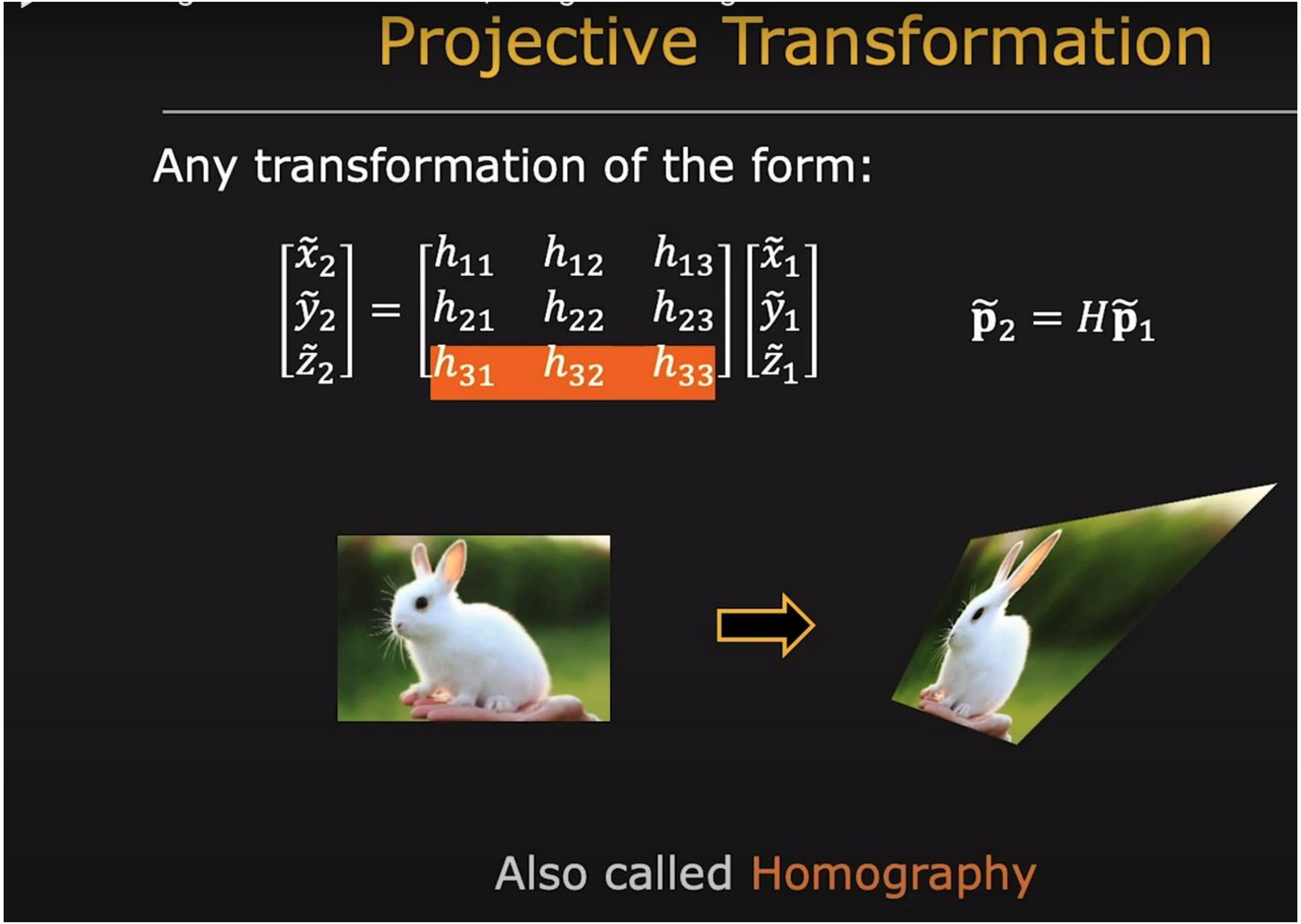


**Fig. 1** First column: images of two consecutive frames overlaid; second column: optical flow (Farnebäck, 2003) between the two frames; third column: optical flow after removing camera motion; last column: trajectories removed due to camera motion in white.



# Projective Transformation

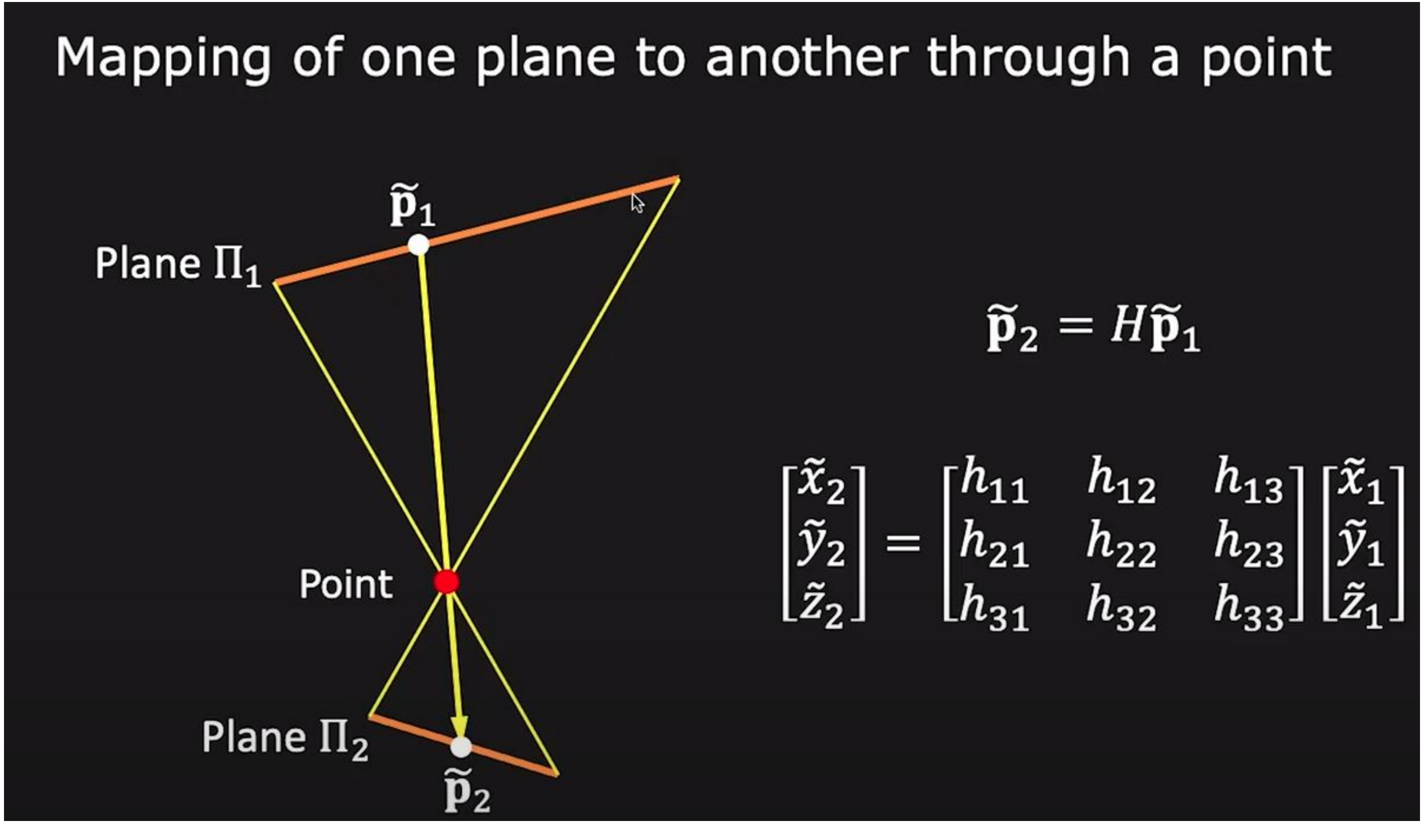
Any transformation of the form:

$$\begin{bmatrix} \tilde{x}_2 \\ \tilde{y}_2 \\ \tilde{z}_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{y}_1 \\ \tilde{z}_1 \end{bmatrix}$$

$$\tilde{\mathbf{p}}_2 = H\tilde{\mathbf{p}}_1$$

Also called Homography

References ;
https://www.youtube.com/watch?v=B8kMB6Hv2eI&list=PL2zRqk16wsdp8KbDfHKvPYNGF2L-zQASc&index=3

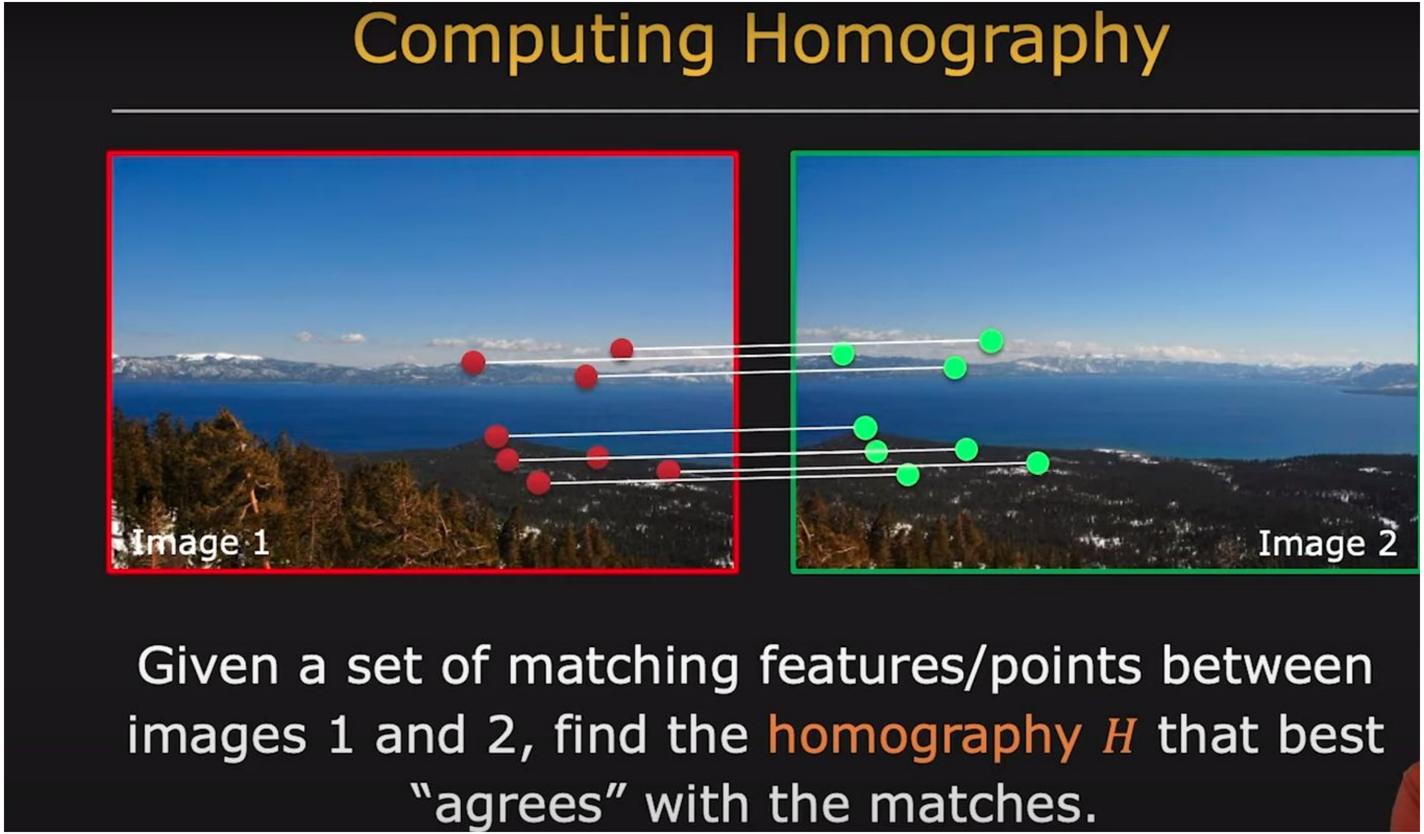


# Mapping of one plane to another through a point

$$\tilde{\mathbf{p}}_2 = H\tilde{\mathbf{p}}_1$$

$$\begin{bmatrix} \tilde{x}_2 \\ \tilde{y}_2 \\ \tilde{z}_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{y}_1 \\ \tilde{z}_1 \end{bmatrix}$$

References ;
https://www.youtube.com/watch?v=B8kMB6Hv2eI&list=PL2zRqk16wsdp8KbDfHKvPYNGF2L-zQASc&index=3



# Computing Homography

Given a set of matching features/points between images 1 and 2, find the homography *H* that best “agrees” with the matches.

References ;
https://www.youtube.com/watch?v=l_qjO4cM74o&list=PL2zRqk16wsdp8KbDfHKvPYNGF2L-zQASc&index=4

Some points on  Dense trajectory features:

- The dense trajectory features approach (Wang et al, 2013a) densely samples feature points for several spatial scales.

- We remove static feature trajectories as they do not contain motion information, and also prune trajectories with sudden large displacements.

- For each trajectory, we compute HOG, HOF and MBH descriptors with exactly the same parameters as in (Wang et al, 2013a).

- Both HOF (Laptev et al, 2008) and MBH (Dalal et al, 2006) measure motion information, and are based on optical flow.

# Procedure for Camera Motion Estimation

- Authors assume that two consecutive frames are related by a homography. Because, global motion between frames is small.

- For, homography estimation we need relation between the frames.For, that we need to consider keypoints for that they used SURF method.Because, SURF features are robust to motion blur.

- Authors also used optical flow algorithm(Farneback, 2003) for sampling motion vectors which provides dense matches between the frames.

  Motion vector is selected for good-features-to-track criterion (Shi and Tomasi, 1994).

The two approaches SURF and  (Shi and Tomasi, 1994) are complementary to each other.

Because, Former focuses on blob-type structure.

The latter focuses on corners and edges.

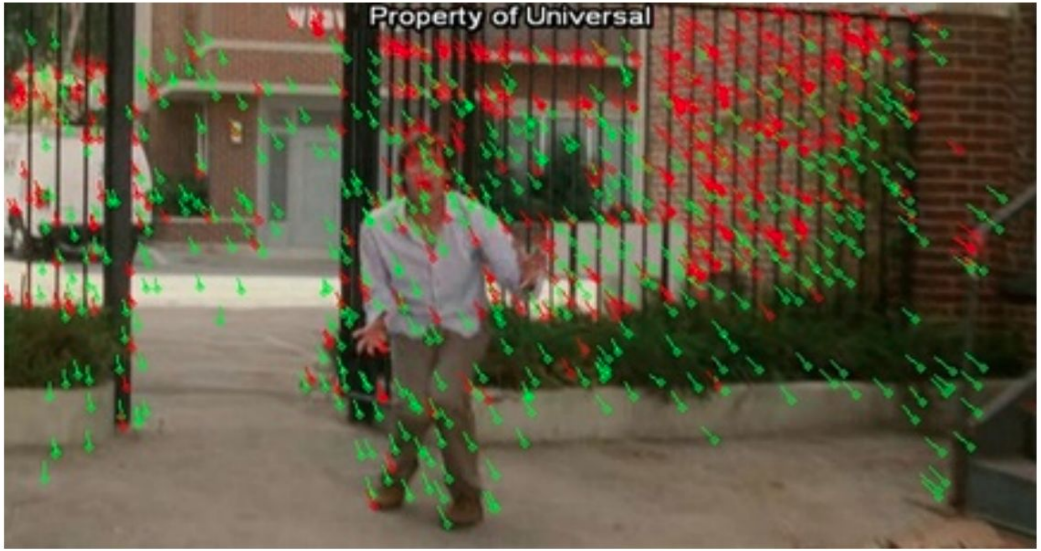


**Fig. 2** Visualization of inlier matches of the estimated homography. Green arrows correspond to SURF descriptor matches, and red ones are from dense optical flow.

# Procedure for Camera Motion Estimation

- Homography estimation is done using random sample consensus method (RANSAC; Fischler and Bolles, 1981).
- Then, they rectify the image using homography to remove the camera motion.

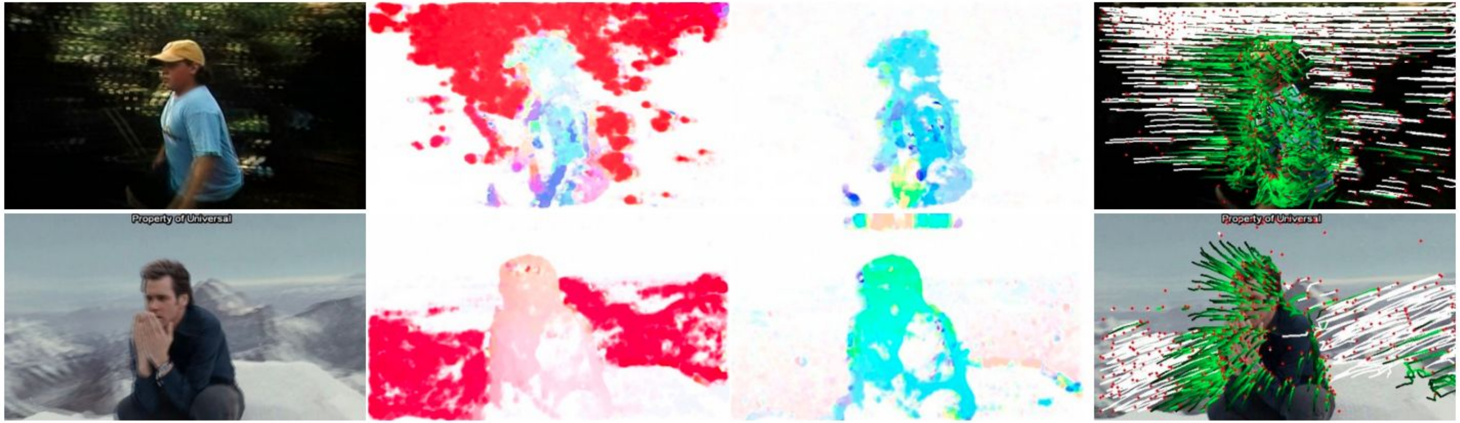


**Fig. 1** First column: images of two consecutive frames overlaid; second column: optical flow (Farnebäck, 2003) between the two frames; third column: optical flow after removing camera motion; last column: trajectories removed due to camera motion in white.

# Advantages of removing camera motion for  trajectory features

- Performance of Motion descriptor like HOF is not degraded significantly which occur in the other case.

- We, can remove trajectories generated by camera motion.By, thresholding displacement vectors of trajectories in warped flow field.

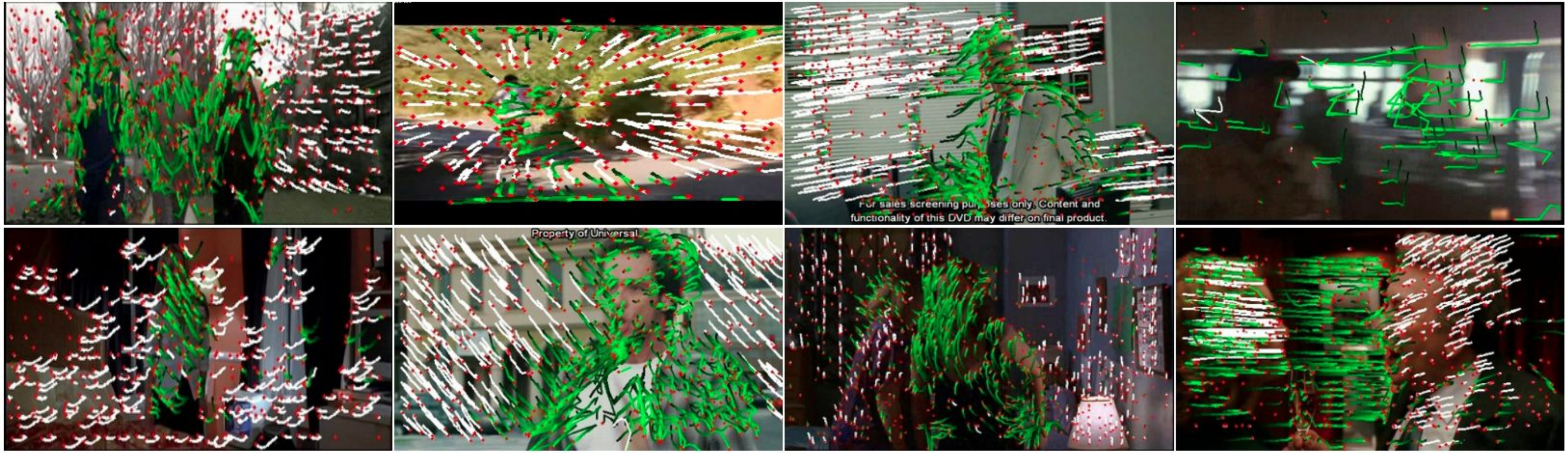

For sales screening purposes only. Content and functionality of this DVD may differ on final product.
Property of Universal

**Fig. 3** Examples of removed trajectories under various camera motions, *e.g.*, pan, zoom, tilt. White trajectories are considered due to camera motion. The red dots are the feature point positions in the current frame. The last column shows two failure cases. The top one is due to severe motion blur. The bottom one fits the homography to the moving humans as they dominate the whole frame.

# Some points on Failure cases

- In, first case there is severe motion blur which leads to SURF and optical flow algorithm both to fail.

- Humans, dominate in the latter case which leads to failure of homography estimation.Because, many features from moving humans become inlier matches.

# Solution to failure of homography estimation

To, remove inconsistencies introduced by the humans:

Authors proposed to use a SOTA human detector so that the detected human present in the frame can be ignored by the feature detection algorithm.

## Some points on human detector

- Detector combines several parts detectors dedicated to different regions of human body like full person, upper-body, face etc.

- Detection threshold is set to 0.1 for high-recall operating point i.e detection more than 0.1 is considered as positive example.

- It may be possible there may be some failure cases in human detector because of complex human body poses, self occlusions, motion blur etc.

# Solution to the failure cases of human detector

- Authors track all the bounding boxes obtained by the human detector.

- Tracking is performed in forward and backward direction for each frame of the video.

- They track each bounding box for utmost 15 frames and stop if there is a 50% overlap with another bounding box.

# Example of human detection

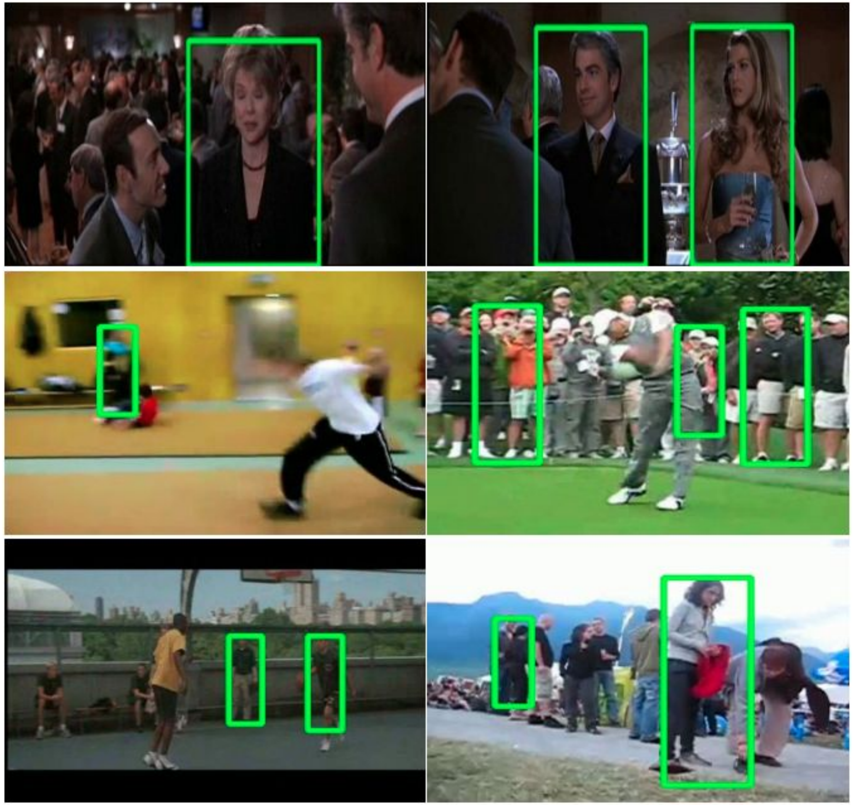


**Fig. 5** Examples of human detection results. The first row is from Hollywood2, whereas the last two rows are from HMDB51. Not all humans are detected correctly as human detection on action datasets is very challenging.

# Procedure to remove camera trajectory via camera motion

- Homography with RANSAC using feature matches is evaluated for each pair of consecutive frames independently to avoid error propagation.
- Warp second frame according to the estimated homography.
- Optical flow is recomputed with the warped frame and first frame.
- HOF and MBH is computed on warped optical flow.
- For, each trajectory magnitude of motion vector is calculated for 15 frames.
- If, maximum magnitude is less than a threshold (set to 1 pixel), the trajectory is considered as camera motion and thus removed.

# Feature encoding

- Fisher vector
- Bag-of-words(BOW)

- Fisher Vector(FV) extends BOW representation as it encodes both first and second order statistics between video descriptor and diagonal covariance GMM(Gaussian mixture model).

Let $x_n \epsilon R^D$ denote nth D-dimensional video descriptor, $q_{nk}$ the soft assignment of $x_n$ to k-th Gaussian, and $\pi_k$, $\mu_k$, $\sigma_k$ are weight, mean and covariance of the k-th Gaussian

The D-dimensional gradients w.r.t the mean and variance of k-th Gaussian are given by:

$$G_{\mu_k} = \sum_{n=1}^{N} q_{nk} \left[ x_n - \mu_k \right] / \sqrt{\sigma_k \pi_k},$$

$$G_{\sigma_k} = \sum_{n=1}^{N} q_{nk} \left[ (x_n - \mu_k)^2 - \sigma_k^2 \right] / \sqrt{2\sigma_k^2 \pi_k}.$$

## Steps to compute FV

- Apply PCA to reduce descriptor dimensionality by 2.
- To,train a GMM with K Gaussians they randomly sample a subset of 1000 x K descriptors.
- After,encoding the descriptors using previous equations they apply power and L2 norm to the final FV.
- A Linear SVM is used for classification.

## Steps to compute BOW histograms

- Soft assignments to the same Gaussians as used in FV.
- Same vocabulary as in FV representation.
- They consider both Linear and RBF-□$^2$ kernel for the SVM classifier.
- For, linear kernel same power and L2 norm as FV, while for RBF L2 norm.

# Weak spatio-temporal location information

According to authors FV and BOW histogram representations are orderless representation.To, add some spatio-temporal information:

They use spatio-temporal pyramid(STP) representation(Laptev et al, 2008), and compute separate BOW or FV over cells in spatio-temporal grids.

# Action Localization

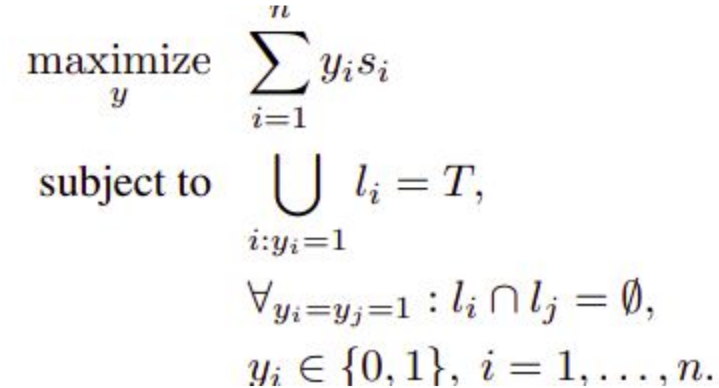
- For, standard NMS authors observe this method has strong tendency to retain short windows.This, is because most characteristic feature of action occur in short sub-sequences.

- To, addres this they tried a variant in which they re-score the segments by multiplying scores with their duration before applying NMS.

# Action Localization

- They also tried another variant with following properties:
  - ❑ Covers the entire video
  - ❑ Does not have overlapping windows
  - ❑ Maximizes the sum of scores of the selected windows.

They express this method as an optimization problem:



$$\begin{aligned} \underset{y}{\text{maximize}} \quad & \sum_{i=1}^{n} y_i s_i \\ \text{subject to} \quad & \bigcup_{i:y_i=1} l_i = T, \\ & \forall_{y_i=y_j=1} : l_i \cap l_j = \emptyset, \\ & y_i \in \{0,1\},\ i = 1, \ldots, n. \end{aligned}$$

$y_1 \ldots y_n$ are boolean variables represent the subset

$s_i$ and $l_i$ denote the score and the i is interval of the window

n is the total number of windows, T is interval that spans the whole video

- The optimal subset is found efficiently by dynamic programming.

- Dynamic programming  Viterbi algorithm is used to get this optimal solution.They refer to this method as DP-NMS.

# Datasets for action recognition

- Hollywood2
- HMDB51
- Olympic Sports
- High Five
- UCF50
- UCF101

# Datasets for action detection

- Coffee and Cigarettes
- DLSB

# Datasets for Event detection

- TRECVID MED 2011

# Conclusion

- Paper improves dense trajectories by removing background trajectories by estimating camera motion.
- They warped the optical flow with a robustly estimated homography approximating the camera motion.
- They explore Fisher vector as an alternative feature encoding approach to bag-of-words histograms with considering the effect of spatio-temporal pyramids,Fisher vectors to encode weak geometric layouts.

# Conclusion

- They found out that action localization improves by simple re-scoring technique before applying NMS.

- There proposed pipeline significantly outperform the SOTA on all three tasks.

- There approach can serve as a general pipeline for various video recognition problems.