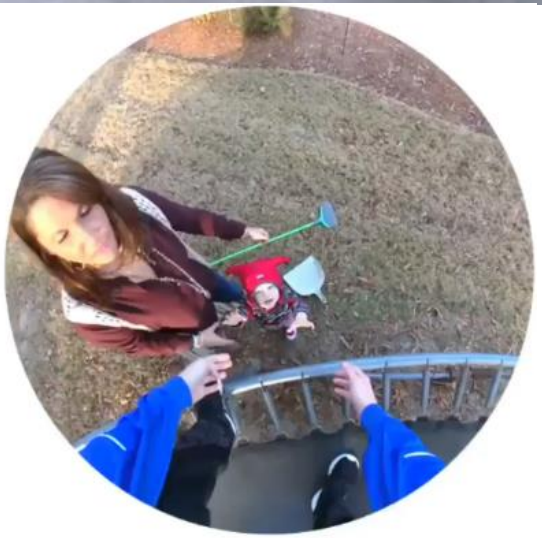
First-Person Activity Recognition with Multimodal Features

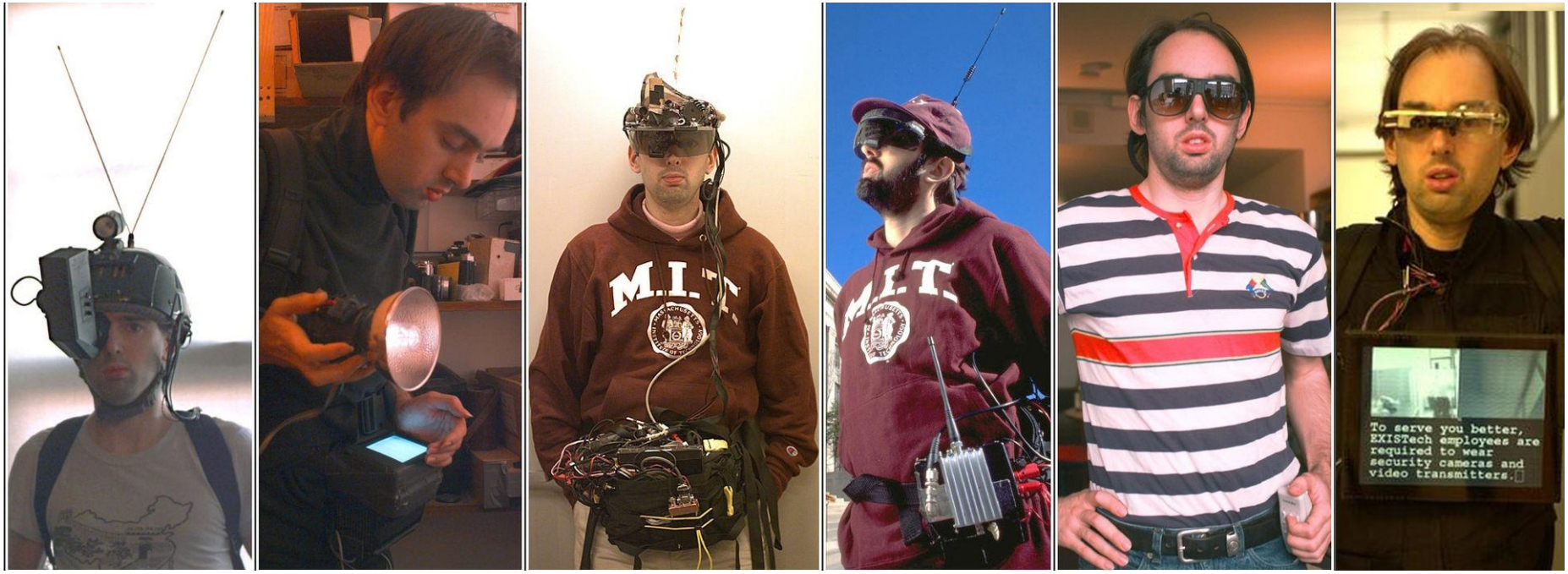
Dr. Alptekin Temizel

Associate Professor
Graduate School of Informatics
Midlle East Technical University (METU)
Visiting Academic
University of Birmingham

First-Person Vision



Wearable Cameras



S. Mann, "'WearCam' (the wearable camera): Personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis," in *2nd Int. Symp. Wearable Comput. Dig. Papers*, Oct. 1998.

Wearable Cameras

Wearable Cameras

- Practical and affordable wearable products became available
  - GoPro (Action camera)
  - Microsoft Hololens (Holographic computer)
  - Snap spectacles (Smart glasses)
  - Google Glass Enterprise Edition (Wearable computer)
- Some discontinued/cancelled
  - Microsoft SenseCam/Vicon Revue/Autographer (Lifelogging camera- still pictures only)
  - Jawbone
  - Google Glass

Wearable Cameras

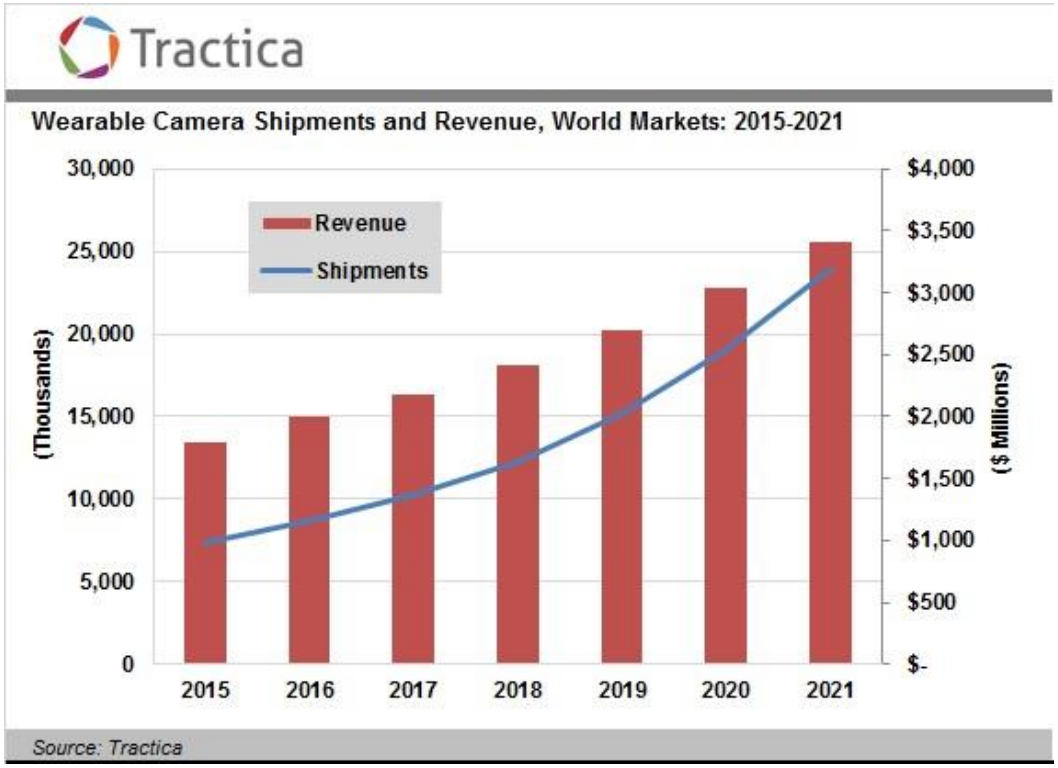
Consumer segments

Established

- Sports and adventure
- Public safety (police officers)

Limited Use

- Lifelogging
- Industrial, enterprise
- Holographic computers



Wearable Cameras

- Commercial products have various types of embedded sensors
  - Video cameras
  - Microphones
  - Eye tracking sensors
  - Accelerometers
  - Gyroscopes
  - Magnetometers
  - Light sensors
  - Proximity sensors
  - Body-heat detectors or temperature sensors

Third-Person Analysis

- The camera points towards the actor(s) involved in the event
- Applications
  - Actions by a person (e.g. walking, running, jumping, dancing)
  - Interactions of multiple persons (e.g. chatting, fighting)
  - Interactions within a group of people (e.g. group formation, identification of emergent leaders)
  - Annotation of sports videos
  - Analysis of crowds / events
  - Face detection, recognition
  - Customer analysis (gender, age-group)

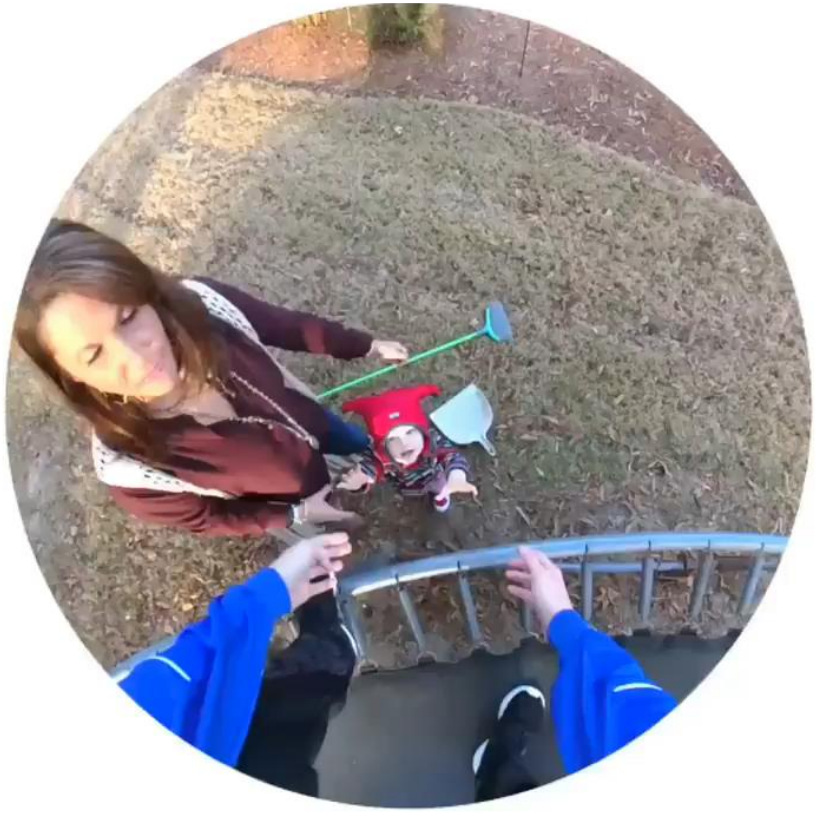
First-Person Analysis

- First-person perspective
  - The observer itself is involved in the events
  - The camera undergoes large amounts of ego-motion with the activity of the user

"BAR Dataset"

Abebe, G., Cavallaro, A. and Parra, X., 2016. Robust multi-dimensional motion features for first-person vision activity recognition. Computer Vision and Image Understanding, 149, pp.229-248.

First Person Vision



Video from a public tweet by @Charliekoehn10
Source: https://twitter.com/Charliekoehn10/status/804901637642944513

First Person Vision

- Observation from a robot's perspective – reactions against a robot

M. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?", IEEE CVPR, 2013.

First Person Vision

"Although we have witnessed impressive progress in several specific applications, our opinion is that the field is only at its beginning"

Guest Editorial Special Issue on Wearable and Ego-Vision Systems for Augmented Experience
G Serra, R Cucchiara, KM Kitani, J Civera - IEEE Transactions on Human-Machine Systems, 2017

INFORMATICS INSTITUTE

multimedia informatics

Advantages

- Captures the main interest points
- Can infer from
  - the ego-motion
  - detected objects
  - changes in illumination and scene
- Multimodality: other sensors are also available

Challenges

Challenges from the point of computer vision

- Moving camera/Egomotion
- High variability of the data
- Changing illumination conditions

Challenges

Challenges from the point of capture and processing

- Privacy issues
- Massive amounts of data
  - Most are uninteresting and repetitive
- Computational challenges
  - Real-time processing on embedded systems
  - Cloud computing
  - Offline processing

Datasets

Dataset Name	Year	Objective	Camera Mount	Data Modalities
CMU-MMAC	2008	Activity Recognition (cooking)	Head + external	Video, audio, motion capture, IMU, accelerometer, light intensity
Intel	2009	Object Manipulation	Shoulder	Video sequence (+segmented objects)
UTEgo	2012	**Activity Recognition**	Head	Video
EDSH	2013	Object Recognition and Tracking	Head	Video (hands visible at all times, indoor/outdoor)
BEOID	2014	**Activity Recognition**	Head	Video, audio, gaze
EGO-GROUP	2014	Social Interaction Detection	Head	Video (indoor/outdoor)
EGO-HPE	2014	Head pose estimation	Head	Video
HUJI EgoSeg	2014	**Activity Recognition**	Head (+Youtube)	Video
JPL	2013	**Activity Recognition**	Head (static)	Video
Dog Centric	2014	**Activity Recognition**	Dog	Video
KrishnaCam	2016	Scene Understanding/ **Activity recognition?**	Head (Google Glass)	Video, GPS position, acceleration, body orientation
SUTD	2016	**Activity Recognition**	Head	Video, audio, accelerometer, gravity, gyroscope, linear acceleration, magnetic field and rotation vector
IAR	2016	**Activity Recognition**	Chest	Video
BAR	2016	**Activity Recognition**	Chest	Video

INFORMATICS INSTITUTE

multimedia informatics

Datasets

- HUJI is currently the largest public dataset
- SUTD is the first dataset containing synchronized egocentric video and sensor data
- Different number of subjects in each database (1 - 30 subjects)
- Some taken indoor, some outdoor, others mixed
- Head-mounted cameras have higher amounts of ego-motion due to head movement
- Chest-mounted cameras have more stable videos but include higher amounts of self-occlusions.
- Cameras attached to dogs exhibit vast ego-motion
- Different types of cameras are used and the quality varies between the datasets

First Person Vision – Activity Recognition

Terminology Summary

Observer type and viewpoint

Type of motion

Objective

Activity Recognition

Activity Recognition

What is the person doing?

F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada and J. Macey. Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, July, 2009.

What is being done to a person?

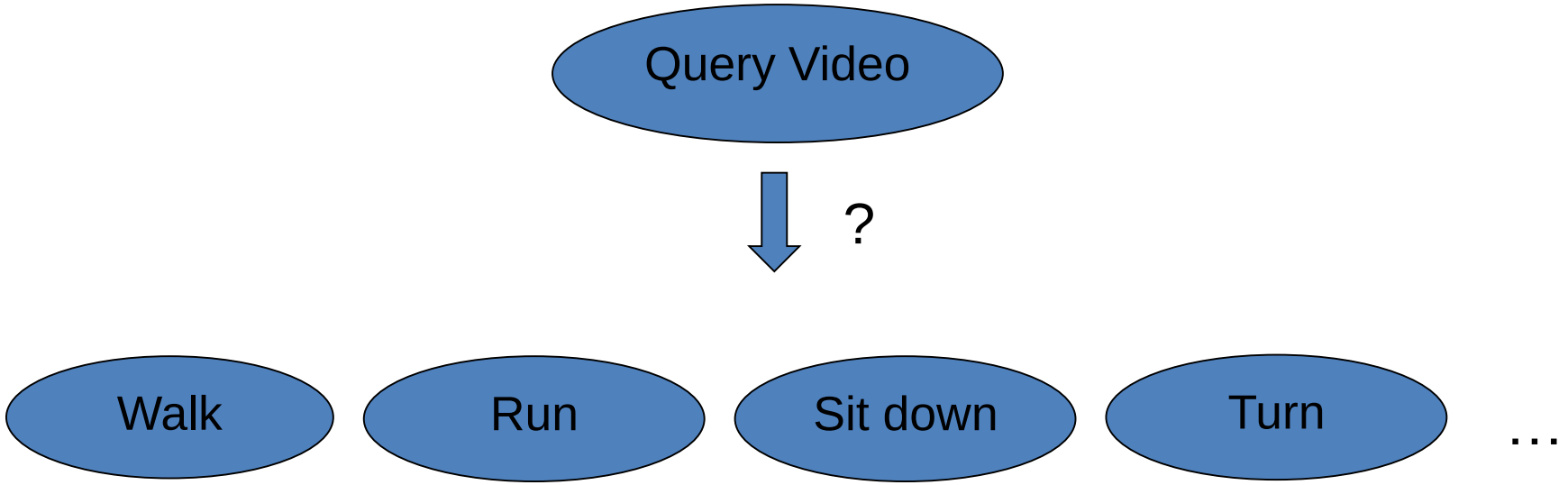
S. Ryoo and L. Matthies, "First-Person Activity Recognition: What Are They Doing to Me?," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013

INFORMATICS INSTITUTE

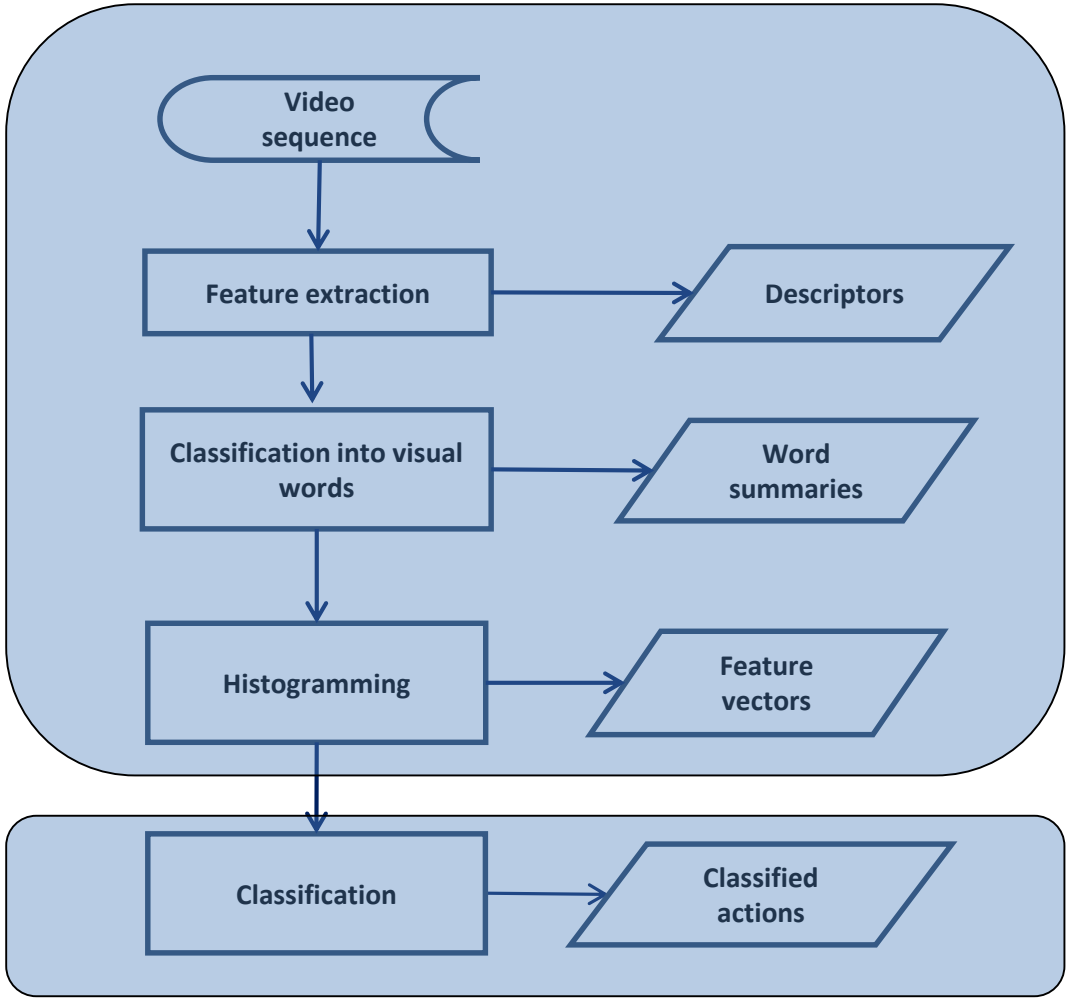
multimedia informatics

Activity Recognition

A supervised learning problem in which the query action class is determined based on a dictionary of labeled action samples.



Activity Recognition



Representation (Model)

Classification

Lower-level Features (Video)

- Local motion descriptors
  - Modelling the other persons/moving objects
- Global motion descriptors
  - Modelling the ego motion

Lower-level Features (Video)

2-D (Intra-Frame) Local motion descriptors*

- Feature Point Descriptors (HOG/SIFT/SURF etc.)

- OverFeat - Convolutional Neural Network (CNN) based
  - Descriptors extracted from the last hidden layer of CNN

*Pooling could be applied to capture temporal variation.

Lower-level Features (Video)

3-D (Inter-Frame) Local motion descriptors

- Cuboids
  - Detect 3D (XYT) interest points
  - Describe 3D spatiotemporal volume by: normalized pixel values, brightness gradient, windowed optical flow
  - Convert into a vector by: flattening, global histogramming or local histogramming

- Space-time interest points (STIP)
  - Detect interest points for a fixed set of multiple spatio-temporal scales
  - Use HOF and HOG to describe patches

Lower-level Features (Video)

Global motion descriptors

- Generic descriptors
  - Histogram of Optical Flow (HoF)
  - Colour histogram
  - Local Binary Patterns (LBP)
  - GIST
  - Log-Covariance
    - 12 dimensional optical flow based motion-related features and intensity-based gradient vectors

Lower-level Features (Video)

Global motion descriptors

- Specifically developed for first person vision
  - Grid optical flow-based features (GOFF) : a set of feature subgroups
    - Motion Magnitude Histogram Feature (MMHF)
    - Motion Direction Histogram Feature (MDHF)
    - Motion Direction Histogram Standard-deviation Feature (MDHSF)
    - Fourier Transform of Motion direction Across Frame (FTMAF)
    - Fourier Transform of grid Motion Per Frame (FTMPF)
  - Virtual Inertial Features (VIF)
    - zero-crossing (ZC)
    - Minimum, maximum, median, energy, kurtosis, mean and standard deviation (4MEKS)
    - Frequency-based feature (FF)

Abebe, G., Cavallaro, A. and Parra, X., 2016. Robust multi-dimensional motion features for first-person vision activity recognition. Computer Vision and Image Understanding, 149, pp.229-248.

Lower-level Features (Other Sensors)

- Accelerometer
- Gyroscope
- Physiological data (e.g. heart rate)
- Audio signal properties
  - Mel-frequency cepstrum
  - Bottleneck features

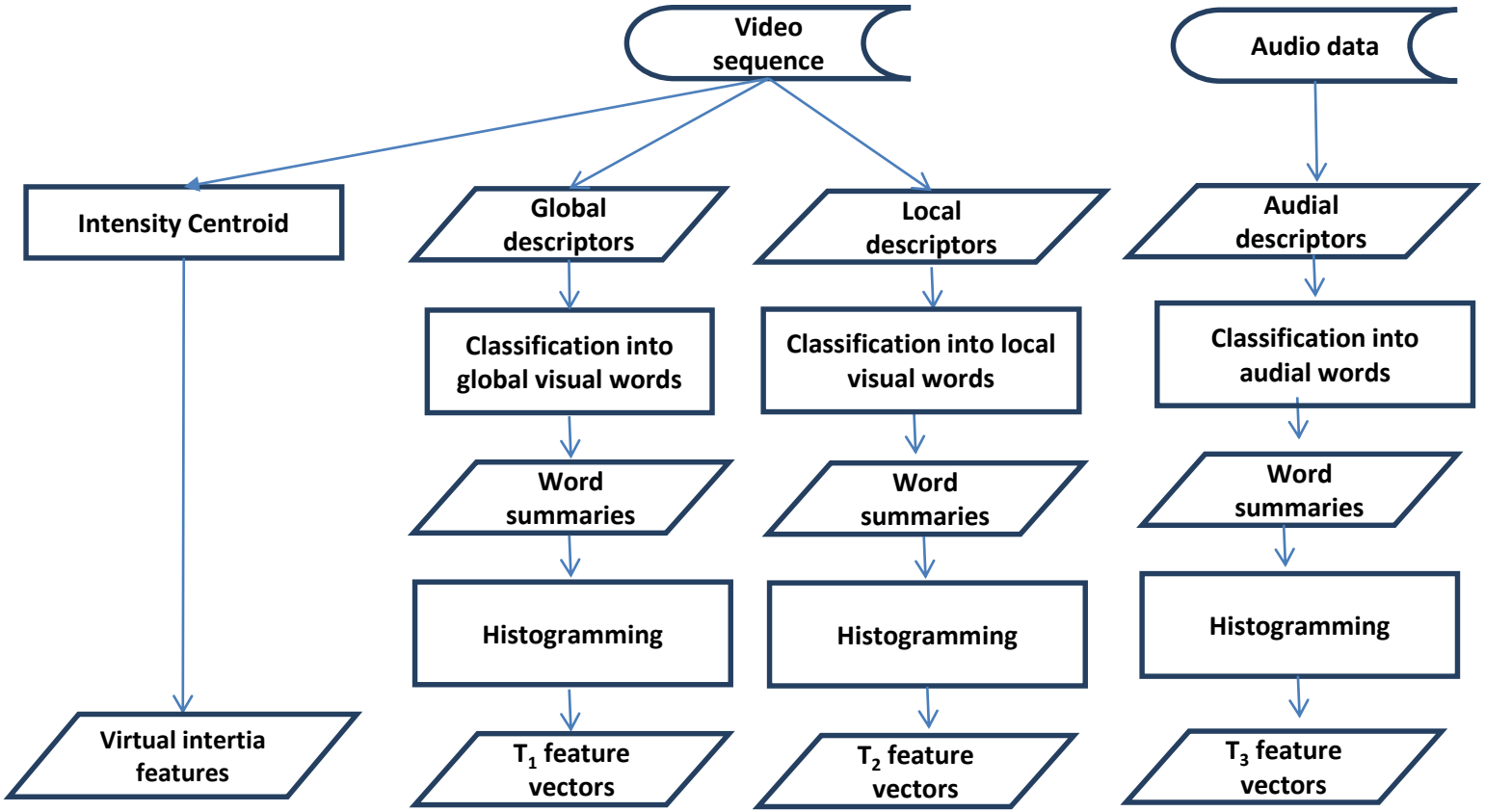
Higher-level Features (Vision)

- Hand segmentation
- Foreground object segmentation
- Car detection
- Person detection
- Detection of interactions between objects

Higher-level Features (Other Sensors)

- Gaze
- Speech activity detection
- Speaker diarisation
- Automatic speech recognition

Activity Recognition



Fusion of Features and Classification

- Use established techniques for classification:
  - SVM
  - Logitboost
  - Logistic Regression
  - KNN
  - Decision Tree
  - HMM
- But we also have:
  - Many features from a single modality
  - Features from different modalities

Kernel Learning

- Kernel learning methods (such as SVM) use a user specified **kernel** (similarity function) over pairs of data points.
- Kernel functions enable operating in a high-dimensional, implicit feature space without computing the coordinates of the data in that space.
- This operation is often computationally cheaper than the explicit computation of the coordinates (**Kernel trick**).
- A linear model can be turned into a non-linear model by applying the kernel trick to the model.

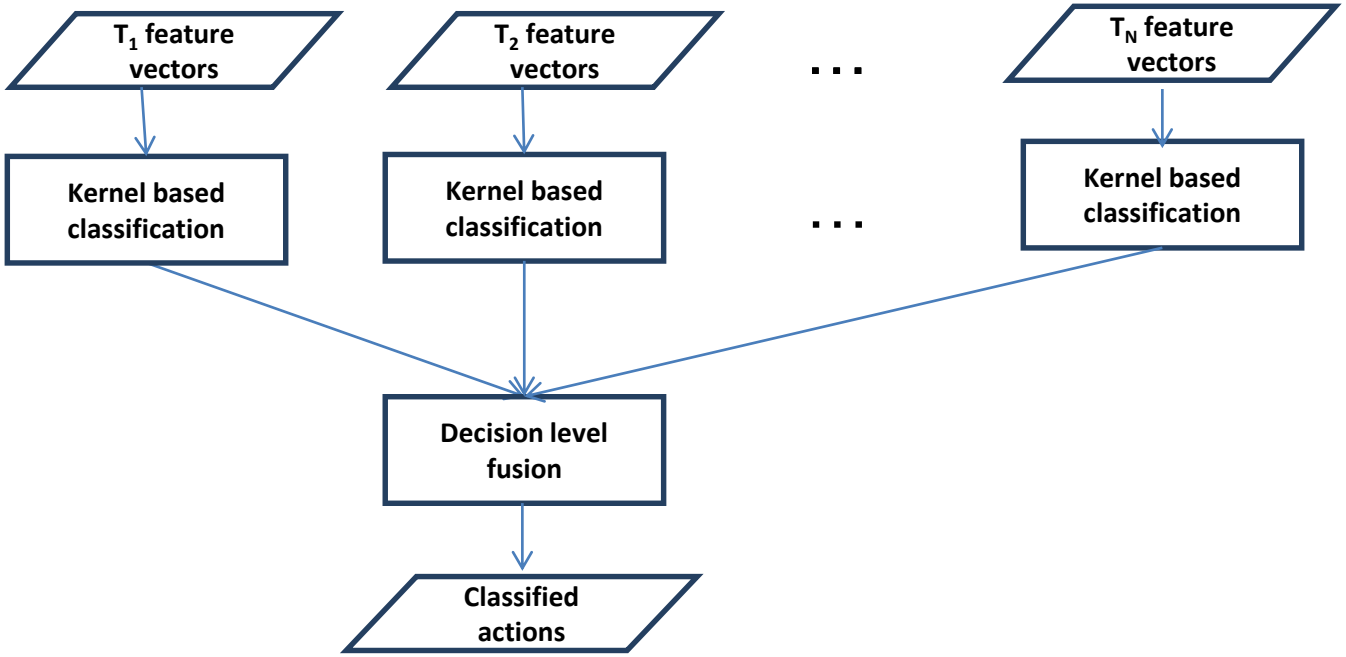
- Different features/modalities require different kernels.

Kernel Learning

General practice in vision applications:

- Assume a predefined parametric kernel, commonly used kernel types
  - Linear
  - Polynomial kernel
  - Radial-Basis Function (RBF)
- Determine the parameters of the kernel function by cross validation
  - Order of the polynomial (Polynomial kernel)
  - Width of the Gaussian (RBF)

Activity Recognition



Multi-Channel Kernels

Multi-channel kernels combines kernels by a pre-set formula :

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\sum_{m=1}^{\mathrm{M}} K_m(\boldsymbol{x}_{i,m}, \boldsymbol{x}_{j,m})}$$

M. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?", IEEE CVPR, 2013.

Multi-Kernel Learning

- Data may come from different sources or might have different representations.
- Multi kernel learning (MKL) method aims to construct an optimal kernel which is a combination of kernels.
- In this way, it is aimed to achieve combining features having different characteristics while reducing bias due to kernel selection.
- Weights to combine kernels for different features are obtained during the learning, facilitating learning and fusion in a single framework.

Multi-Kernel Learning

- Kernels could be variations of kernels for a single feature (families of kernels)
  - In general practice, a single kernel type/parameter is selected in an ad-hoc way
  - MKL facilitates using a range of kernel parameters and optimizes their combination
  - For example: Gaussian kernels with width parameters [0.5 1 2 5 7 10 12 15 17 20]
- Kernels could be for different features
  - MKL performs a data-driven feature learning/selection/weighting
  - For example: In egocentric vision, these could be global and local features

We could benefit from both together!

Multi-Kernel Learning

Add an extra parameter c_m to the minimization problem of the learning algorithm

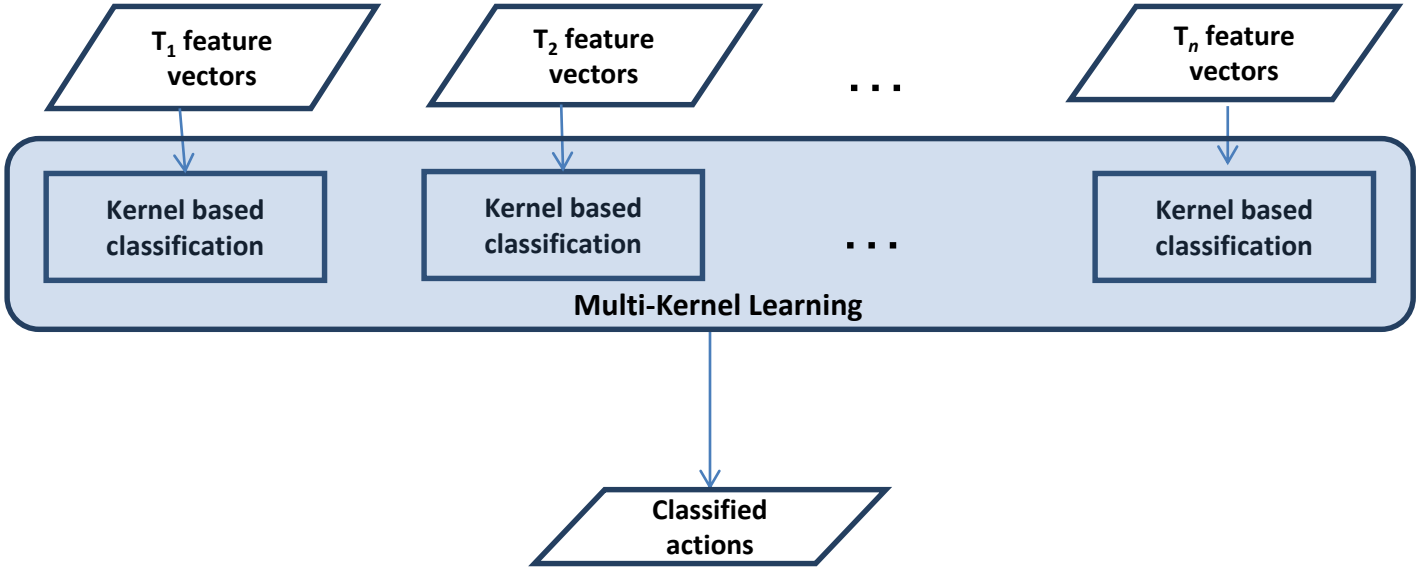
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{\mathrm{M}} c_m K_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

where $c_m \geq 0$ and $\sum_{m=1}^{M} c_m = 1$

c_m are learned in conjunction with a predictor

Solve using standard optimization methods

Activity Recognition



Multi-Kernel Learning Optimization Methods and Tools

- MKL-SD: Solved by a semidefinite program (SDP) which is then resolved by applying some existing optimization techniques [1].
- MKL-SILP: min-max optimization to find the saddle-point solution by solving a semi-infinite linear program (SILP) [2].
- MKL-Level: Solved by an extended level optimization method [3].
- MKL-Hessian: Solved by a second order Newton update optimization method [4].
- Lp-MKL: Generalizes the regular ‘1-norm MKL method to arbitrary ‘p-norm MKL [5].
- SimpleMKL: Iteratively determine the combination of kernels by a gradient descent wrapping a standard SVM solver [6]
- LMKL – Localised multiple kernel learning [7] – Non--linear

Multi-Kernel Learning Optimization Methods and Tools

[1] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan, "Learning the Kernel Matrix with Semidefinite Programming," J. Machine Learning Research, vol. 5, pp. 27-72, 2004.

[2] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf, "Large Scale Multiple Kernel Learning," J. Machine Learning Research, vol. 7, pp. 1531-1565, 2006.

[3] Z. Xu, R. Jin, I. King, and M.R. Lyu, "An Extended Level Method for Efficient Multiple Kernel Learning," Proc. Neural Information Processing Systems (NIPS), 2008.

[4] O. Chapelle and A. Rakotomamonjy, "Second Order Optimization

of Kernel Parameters," Proc. Advances in Neural Information Processing Systems (NIPS) Workshop, 2008.

[5] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Muller, and A. Zien, "Efficient and Accurate 'p-Norm Multiple Kernel Learning," Proc. Neural Information Processing Systems (NIPS), pp. 997-1005, 2009.

[6] A. Rakotomamonjy, F.R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," J. Machine Learning Research, vol. 11, pp. 2491-2521, 2008.

[7] Gönen, M. and Alpaydin, E., 2008, July. Localized multiple kernel learning. ACM *Int. Conf. Machine Learning (ICML)*, pp. 352-359, 2008

INFORMATICS INSTITUTE

multimedia informatics

Multi-Kernel Learning Limitations

- Requires computationally complex optimization
- The final classifier is a single kernel-based classifier which is based on a linear combination of multiple kernels (except [7])

Results

Classification Accuracies

	Accuracy (%)	
Approaches	DogC dataset	JPL dataset
Ryoo et al.	60.5	84.4
Abebe et al. (GOFF + VIF)	61.0	86.0
Multi-Kernel Learning		
HOF + Cuboid	64.9	86.1
HOF + Cuboid + Log–C	64.8	85.7
GOFF + VIF +Log-C	**65.0**	**93.1**

INFORMATICS INSTITUTE

multimedia informatics

Results

Classification Accuracies for SUTD Dataset

	Accuracy (%)	
	SVM	MKL
VIF	26	26
VIF + Log-C	42	41
GOFF + VIF + Log-C	61	62
GOFF + VIF + Log-C + Audio	**64**	**69**

Conclusions

- First-person vision based analysis is getting more popular
- An important problem is the designing of the features
- The other important problem is the selection/weighting of the features (and kernels)
- Can utilize combinations of modalities for better analysis
- Combining the features bring new challenges that needs to be addressed- Multi-Kernel Learning (MKL) and Boosted MKL are prominent candidates to deal with heterogeneous data
- Accounting for the true geometry of the data is desirable – (manifolds?)
- Performances could be improved by effective use of temporal information / sub-events for interaction-level activities

Thanks!
Questions?