Utilizing Multiple Instance Learning for Computer Vision Tasks

Fadime Şener

advised by

Assist. Prof. Dr. Pınar Duygulu Şahin

Assist. Prof. Dr. Nazlı İkizler Cinbiş

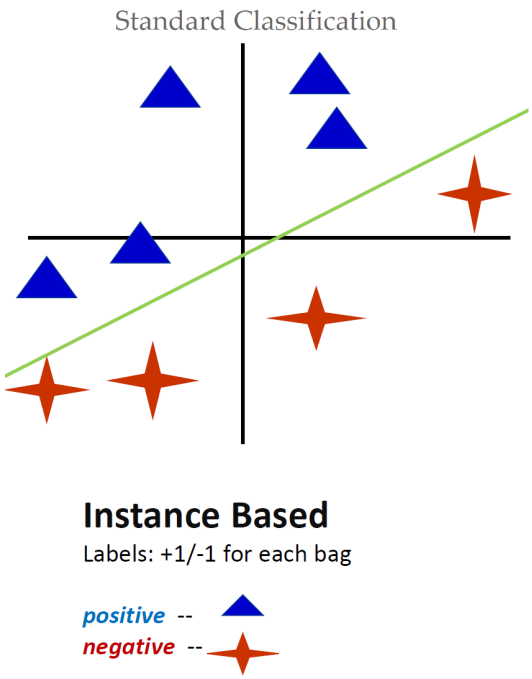
July 8, 2013

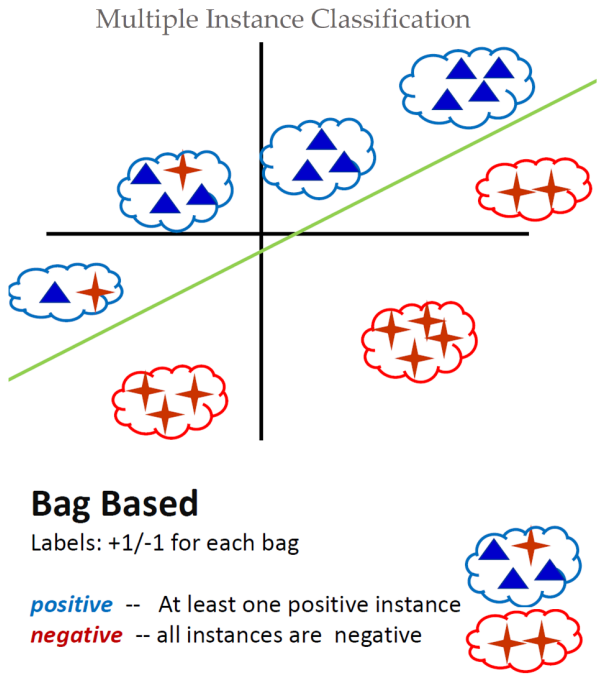
Motivation

- The Multiple Instance Learning (MIL)
  - Arises to be useful in many application domains,
  - Large applicability to problems in computer vision;
    - object recognition, detection, tracking, image classification, scene classification.
  - Particularly suitable for computer vision problems due to the **dificulty of obtaining manual labeling**.
- Three different computer vision problems;
  - Image retrieval and re-ranking,
  - Recognizing actions from still images,
  - Recognition of human interactions in videos.

Multiple Instance Learning

- Multiple Instance Learning operates over bags of instances.
- This form of learning is referred as "weakly supervised"



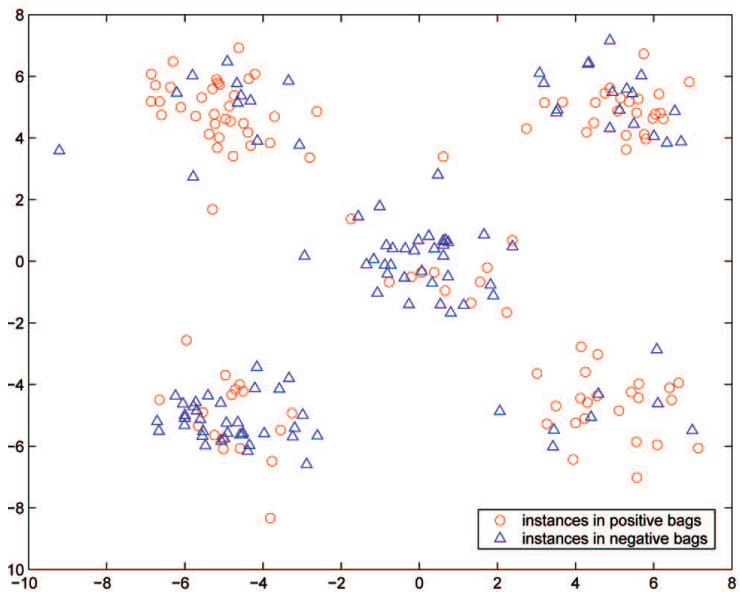


Related Work

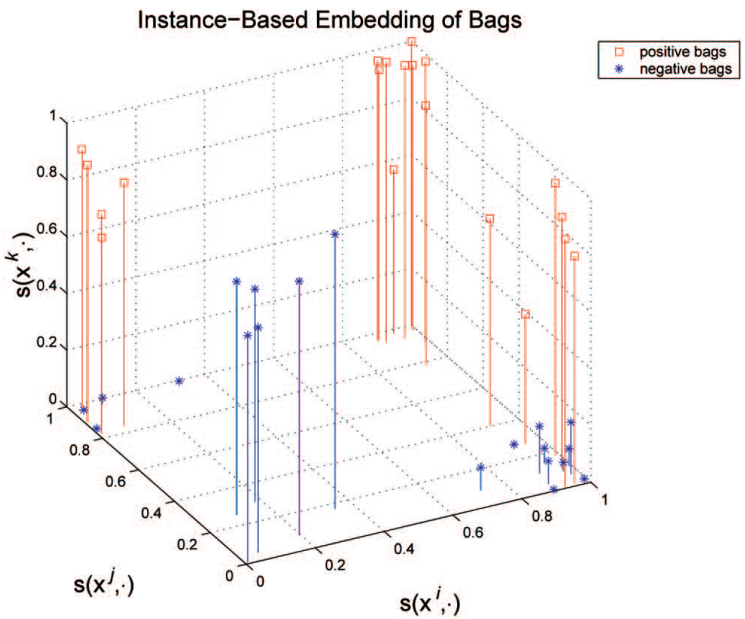
- First used solved problem of drug activity prediction **[Dietterich 1997]**.
- Diverse Density **[Maron NIPS'98]**, Expectation-maximization with Diverse Density - **EM-DD [Zhang NIPS'01]**
- Standard supervised learning techniques used to solve : SVM **MI-SVM** and **mi-SVM [Andrews NIPS'03]** , k-nearest neighbor **[Wang ICML'00]** , boosting **[Andrews NIPS'04]**, neural networks **[Ramon ICML'00]**, decision trees **[Blockeel ICML'05]**.
- Graphical models used ; **[Deselaers ICML'10] [Leistner ECCV'10]**
- **mi-Graph** and **MI-Graph** by treating instances as non-i.i.d. samples **[Zhou ICML'09]**
- Alternative MIL assumptions **DD-SVM [Chen 2004] MILES [Chen PAMI'06]** ,**MILIS [Fu PAMI'11]** .
- MIL has been used in a wide range of applications, such as object recognition and detection **[Viola NIPS'06] [Dollár ECCV'08]** , tracking **[Babenko CVPR'09] [Zeisl BMVC'10]** , image classification **[Zha CVPR'08] [Yakhnenko BMVC'11]**, scene classification **[Maron ICML'98]** and more.

MILES : Multiple-Instance Learning via Embedded Instance Selection

- We use Multiple Instance Learning with Instance Selection (MILES) algorithm **[1]**
- Visualization of instances and bags
  - Instances on a two-dimensional plane
  - Bags embedded in a features space



(a)



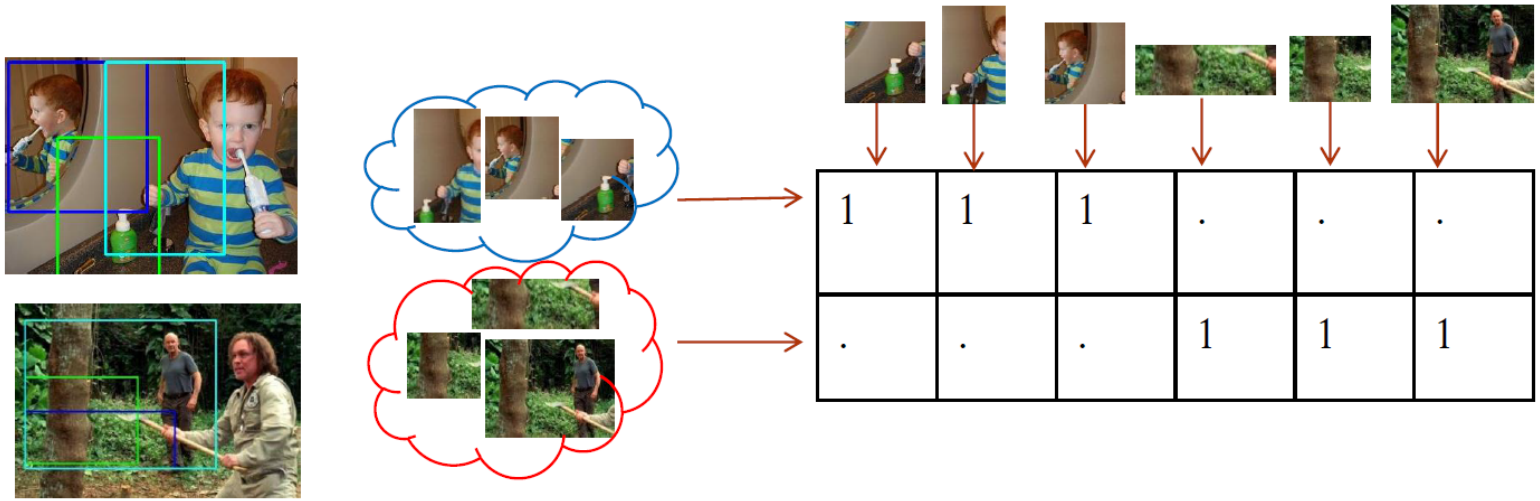
(b)

[1] Chen, Y., Bi, J. and Wang, J. Z., MILES: Multiple-Instance Learning via Embedded Instance Selection , (graphics taken)

MILES : Multiple-Instance Learning via Embedded Instance Selection

- MILES works by embedding the original feature space x, to the instance domain $\mathbf{m}(B)$.
- The similarity between bag $\mathbf{B}_i$ and concept c_l is defined as

$$s(c_l, \mathbf{B}_i) = \max_j \exp\left(-\frac{D(x_{ij}, c_l)}{\sigma}\right), \tag{1}$$



Ensemble of Multiple Instance Classifiers for Image Re-ranking

- Problem Definition
- Related Work
- Our Approach
  - Constructing Candidate Bags
    - Fixed-size
    - Dynamic-Size
    - Sliding Window
  - Evaluation of the bag-size and bag construction approaches
  - Ensemble of MIL Classifiers
  - Comparison to state-of-the art

Problem Definition

- Image retrieval studies
  - content-based image retrieval
  - text-based image retrieval.
- Text-based image search may perform poorly
  - such as irrelevant or incomplete text surrounding the images, polysemy or synonymy of textual descriptions, and so on

- Problem of re-ranking of images returned by text based search engines by visual content analysis

Related Work

- pLSA-based methods**[Fergus ICCV'05]**, LDA-based model **[Fritz CVPR'08]**.
- Reranking mechanism based on spectral filtering and graph based ranking **[Liu CVPR'11]**, PageRank algorithm**[Jing TPAMI'08]**
- Textual features has been explored improving the image reranking **[Berg CVPR'06]**, **[Grangier TPAME'08]**.
- Multimodal features such as text, metadata and visual features together used **[Schroff ICCV'07]**.
- Utilizing the textual and visual features **[Geng CVPR'10]**
- Multiple instance learning to image reranking **[Li ICCV'11]** , **[Duan 2011]** .

Our Approach

- Re-rank images by visual content analysis
  - Automatically create **bags**
    - Fixed-size
    - Dynamic-Size
    - Sliding Window
  - Operate on multiple sized candidate bags and train classifiers using each of the constructed set of bags.
  - Utilize these automatically constructed bags by ensembles of Multiple Instance Learning(MIL) classifiers
  - Re-ranked the images according to the classiffication responses.

Constructing Candidate Bags - Fixed-size

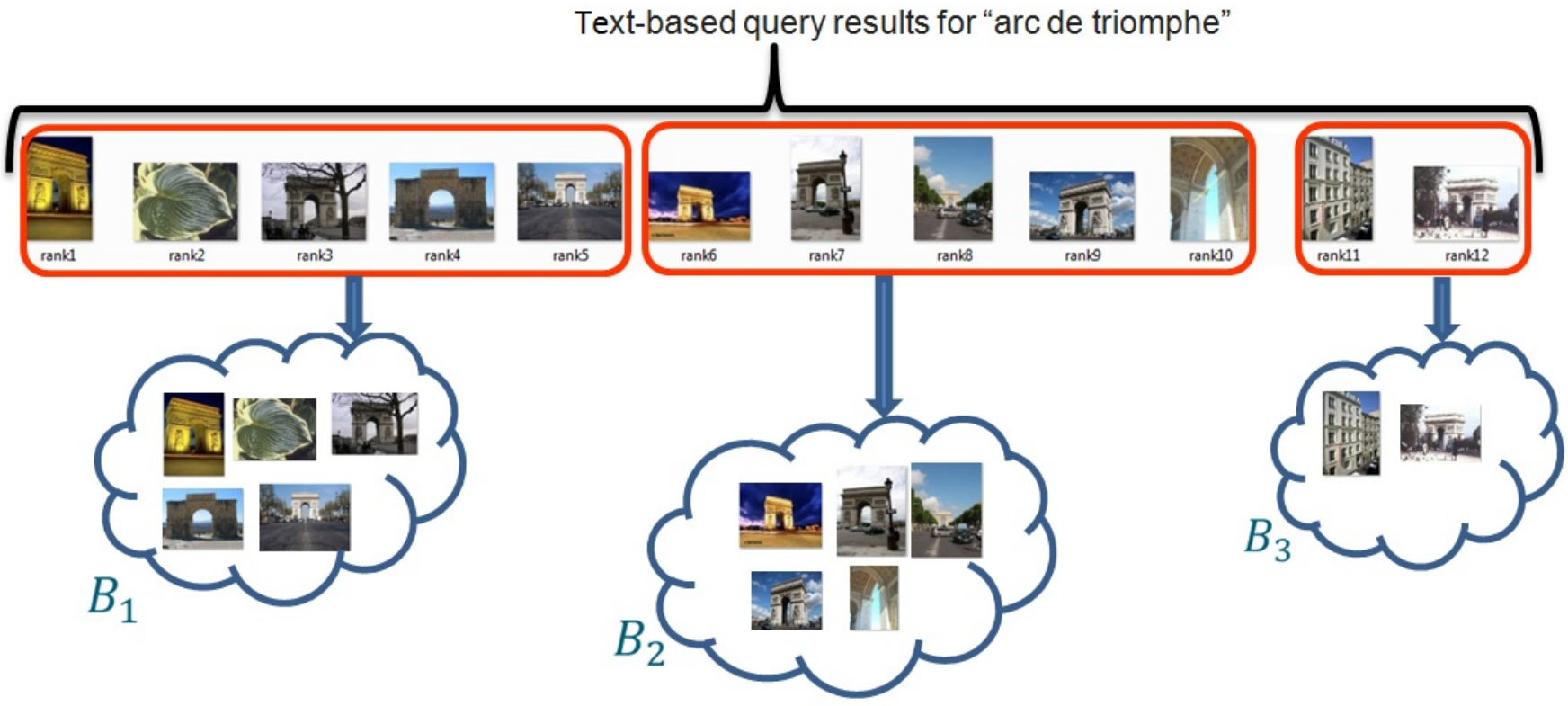


Figure : Formation of fixed-size bags from the retrieved images. In this example $k = 5$ images form individual instances of a single bag, based on the text-based retrieval order. These bags are then fed into multiple instance classifiers as positive bags.

Constructing Candidate Bags - Dynamic-Size

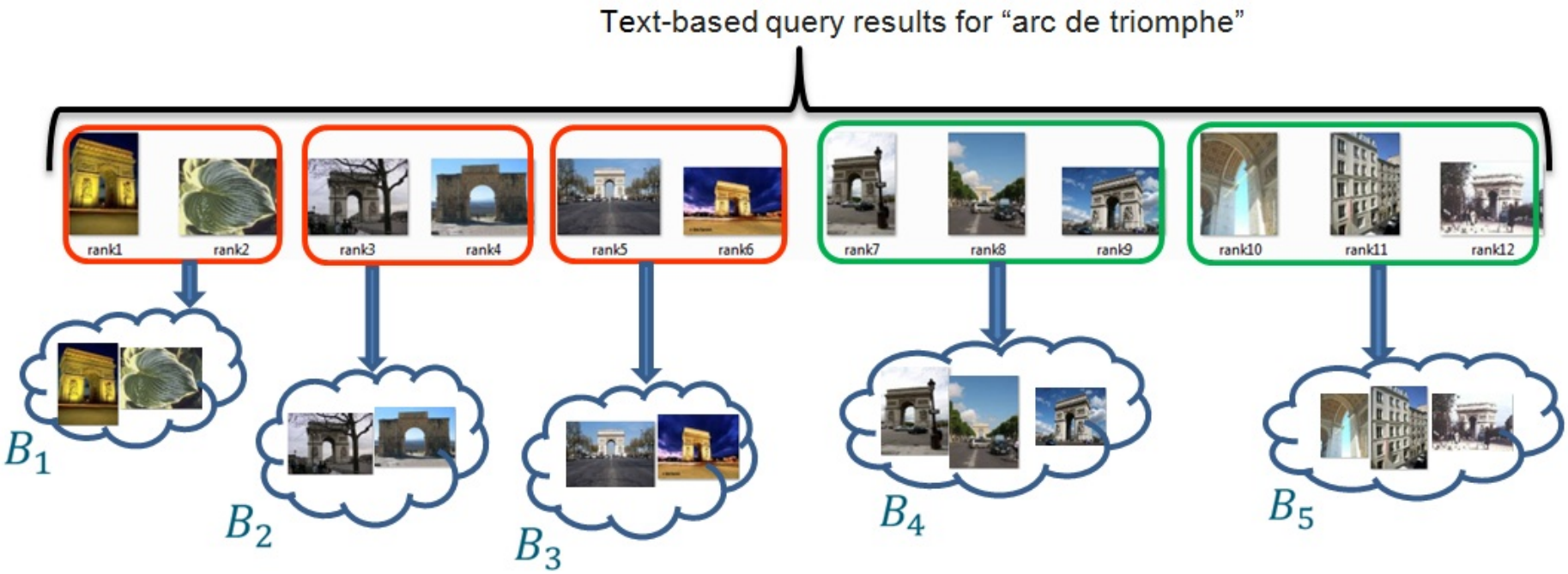


Figure : Formation of dynamic-size bags from the retrieved images. For the images that returned earlier in the list, smaller bags are formed, and for the images that return later in the list, larger bags are formed. In this example, the initial k is 2 and then, for the lower ranks of the text-based retrieval order k value is incremented by 1 and larger bags are formed.

Constructing Candidate Bags - Sliding Window

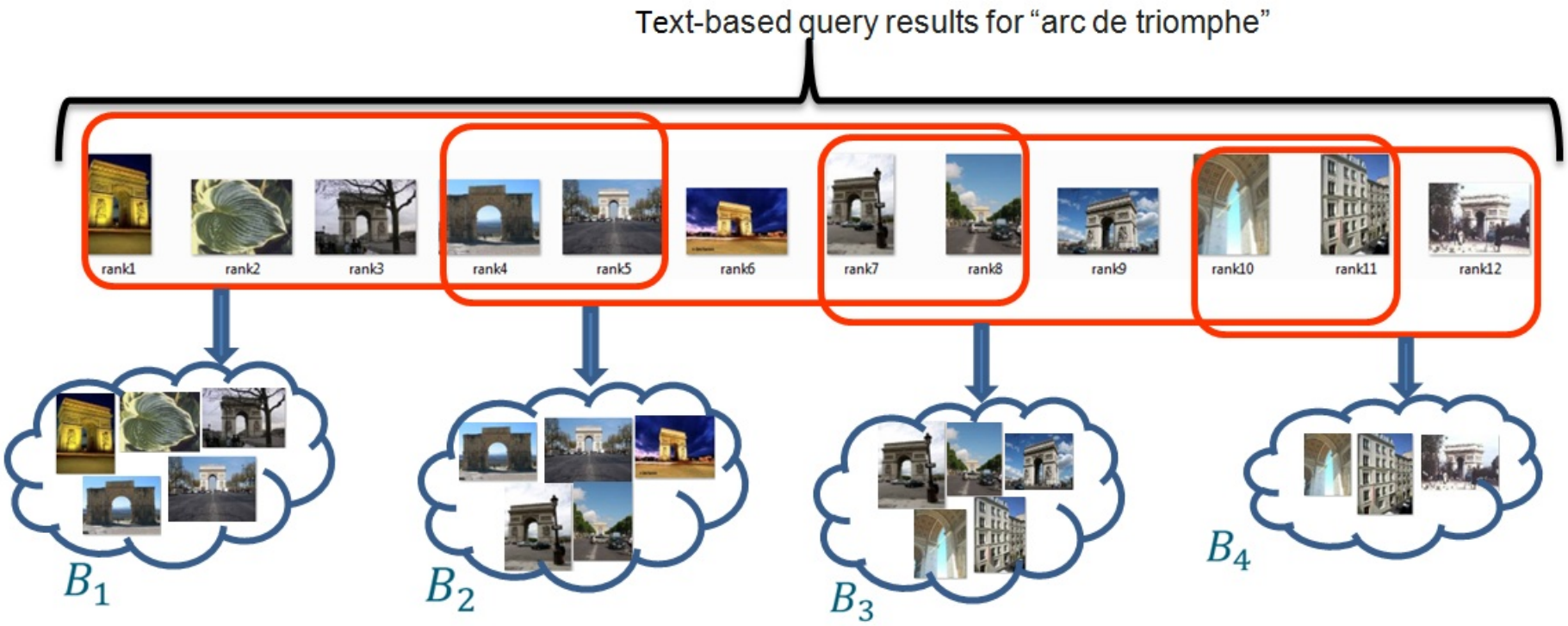
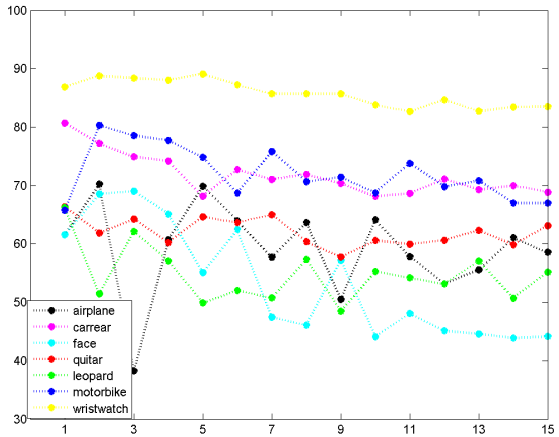


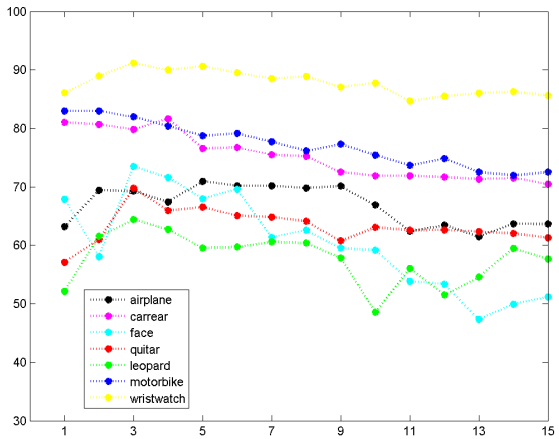
Figure : Sliding window approach for formation of fixed-size bags from the retrieved images. Here k is fixed ($k = 5$) and step size $M = M = ceil(k/2)$. Sliding window approach generates multiple overlapping bags and provides a dense sampling of the possible bag candidates for MI learning.

Evaluation of the bag-size and bag construction approaches

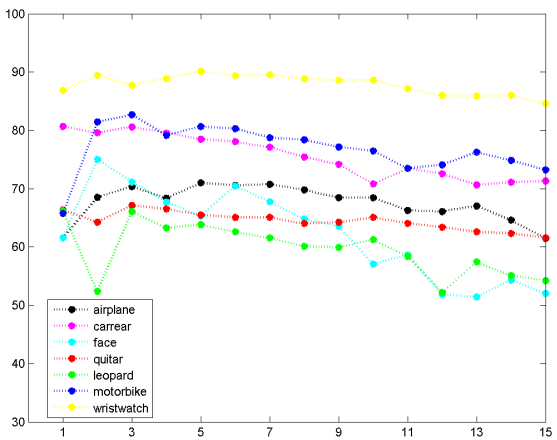
- Effect of choosing different bag sizes k (average precision(AP))



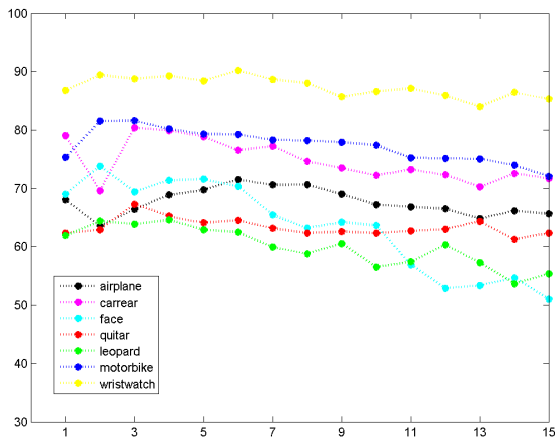
(a) Fixed-size



(b) Dynamic-size



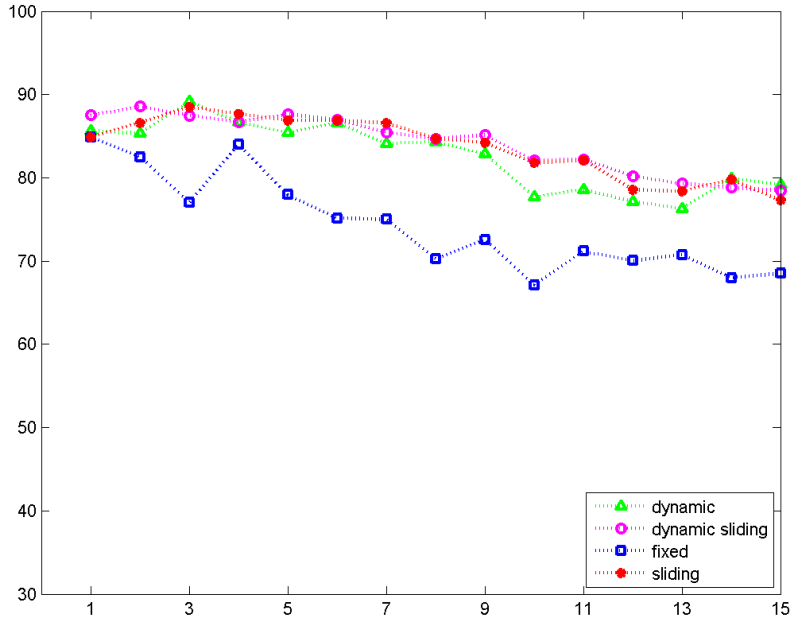
(c) Sliding window

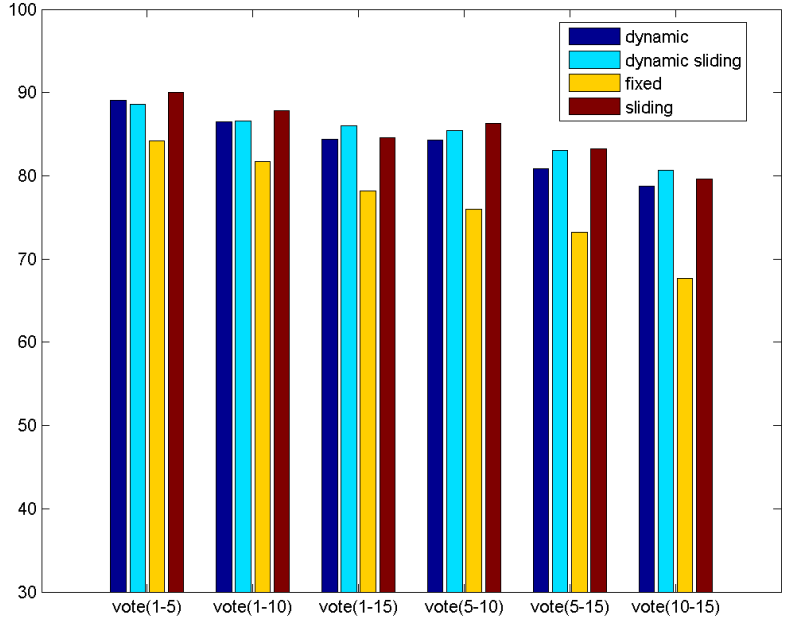


(d) Dynamic-Sliding

Ensemble of MIL Classifiers

- No information on the positivity of the retrieved images. We generate multiple candidate bags with varying k, and train classifiers.
- Left Figure : Mean performance of MI-bag construction methods with respect to changing bag size k.
- Right Figure : Ensemble of MI classifiers with different bag sizes k and different bag construction. vote(k_1,k_2) shows that $k \in k_1 \ldots k_2$.





Comparison to state-of-the art

- Comparison to state-of-the art on Google dataset **[1]**. In this table, precisions (%) at 15% recall are reported

	airplane	car_rear	face	guitar	leopard	mbike	wrwatch	mean
Google	50	41	19	31	41	46	70	43
[41]	35	-	-	29	50	63	93	54
LogReg [2]	65	55	72	28	44	49	79	56
TSI-pLSA [5]	57	77	82	50	59	72	88	69
WsMIL [72]	100	81	57	52	66	79	95	75.7
SF+MRank [38]	86	100	75	58	63	79	100	80
PMIL [43]	100	75.3	89.9	82.7	86.2	76.6	95.7	86.6
LDA [35]	100	83	100	91	65	97	100	91
Ours	100	100	97.5	82.7	75.5	97.1	100	**93.3**

[1]R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search.," in Proceedings of the 10th International Conference on Computer Vision, Beijing, 2005.

Comparison to state-of-the art

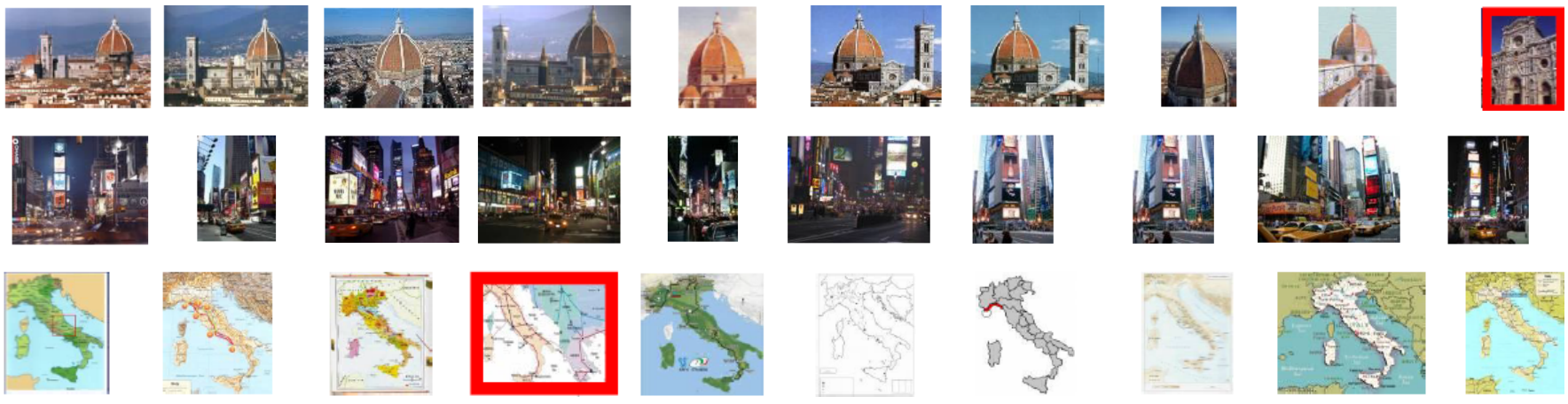
- Comparisons to state-of-the art on Web Queries **[1]** with respect to the Mean average precisions (MAP)

Method	MAP
Search Engine	56.99
[1] (visual only)	64.9
[1] (visual+textual)	67.3
BLVS**[2]**	67.0
SpecFilter+MRank**[3]**	**73.76**
Ours	71.08

[1] J. Krapac, M. Allan, J. Verbeek, and F. Jurie, "Improving web-image search results using query-relative classifiers," in IEEE Conference on Computer Vision Pattern Recognition (CVPR '10),
[2] N. Morioka and J. Wang, "Robust visual reranking via sparsity and ranking constraints," in Proceedings of the 19th ACM international conference on Multimedia,2011.
[3] W. Liu, Y.-G. Jiang, J. Luo, and S.-F. Chang, "Noise resistant graph ranking for improved web image search," in IEEE Conference on Computer Vision Pattern Recognition (CVPR), 2011.

Qualitative Results

- Examples of the retrieval order obtained by our method. Top 10 images for each query are shown.



- Cases of failure. Top 10 images for each query are shown.

Recognizing Actions in Still Images using Multiple Instance Learning

- Problem Definition
- Related Work
- Our Approach
  - Multiple Instance Learning for Candidate Object Regions
  - Additional Features
  - Comparison to state-of-the art
  - Most contributing concept instances

Problem Definition

- Problem of Recognizing Actions in Still Images using Multiple Instance Learning
- Videos : motion and appearance
- Still images convey the action information via ;
  - the pose of the person and the surrounding object/scene context.

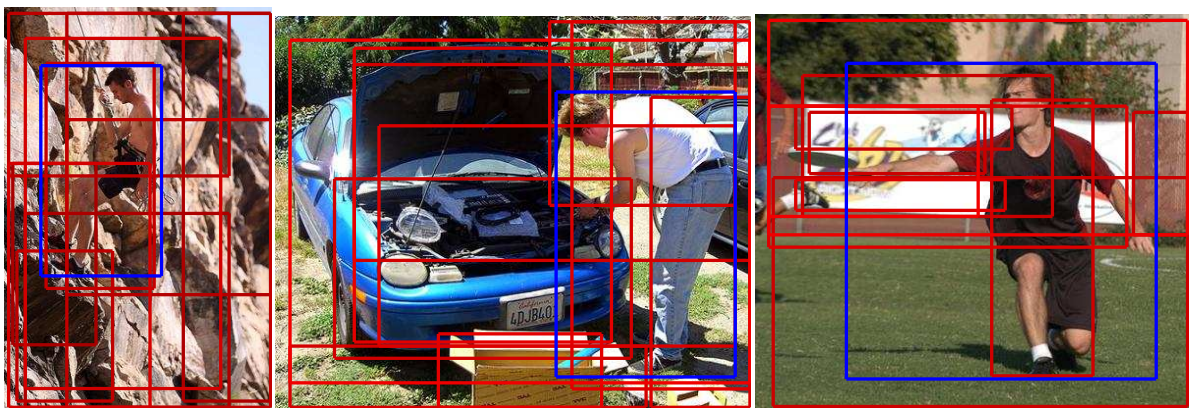


Related Work

- Action recognition in videos has large amount of work
- Action recognition in still images is less studied
  - Deformable template matching for computing the distance between human poses. **[Wang CVPR'06]**
  - Pose primitives are learnt from non-cluttered videos and applied to images for finding the closest pose **[Thurau CVPR'08]**
  - Pose models are learnt from action images and those models are applied to classify actions in videos. **[Ikizler-Cinbis ICCV'09]**
  - Relationship between poses and objects ; grouplet features **[Yao CVPR'10]**, **[Desai 2010]**, **[Gupta PAMI'09]**, **[Delaitre NIPS'11]**
  - Use of explicit object detectors for aiding action recognition in still images. **[Yao ICCV'11]**
  - Weakly supervised learning of human-object interactions **[Prest TPAMI'12]**

Our Approach

- Objects are particularly important.
- We extract several candidate object regions using objectness measure **[1]**.
- We use Multiple Instance Learning(MIL) framework over candidate regions.

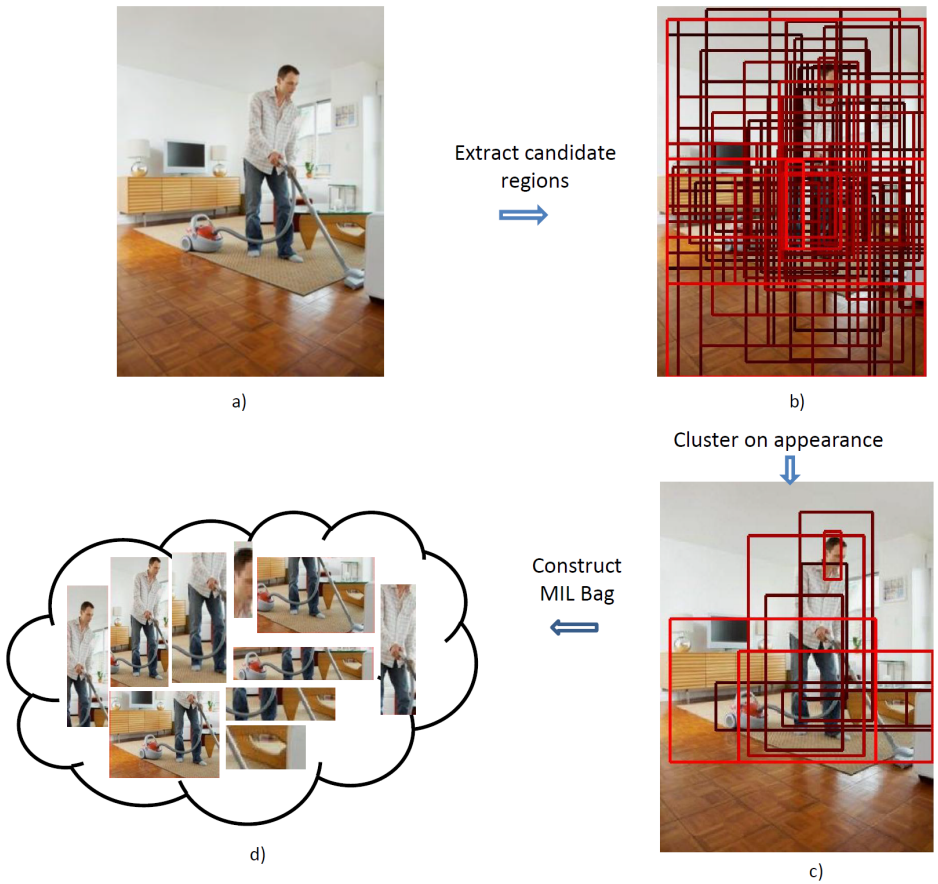


- Besides we evaluate various features

[1] Alexe, B., Deselares, T. and Ferrari, V. , What is an object?

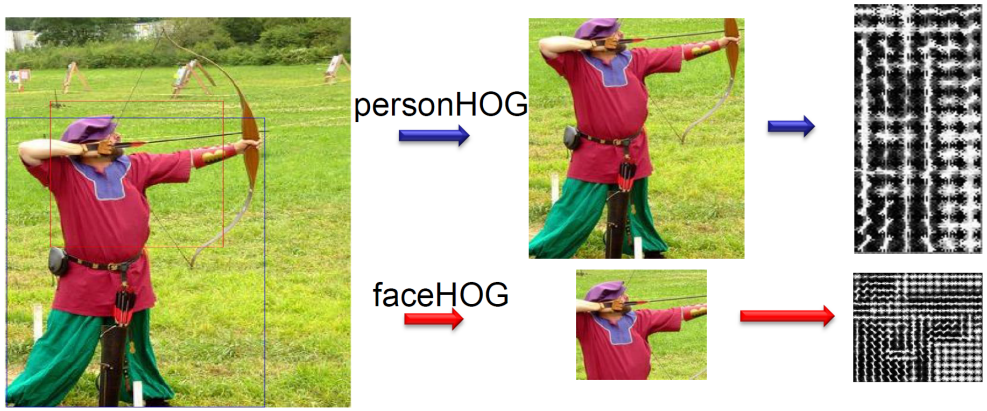
Multiple Instance Learning for Candidate Object Regions

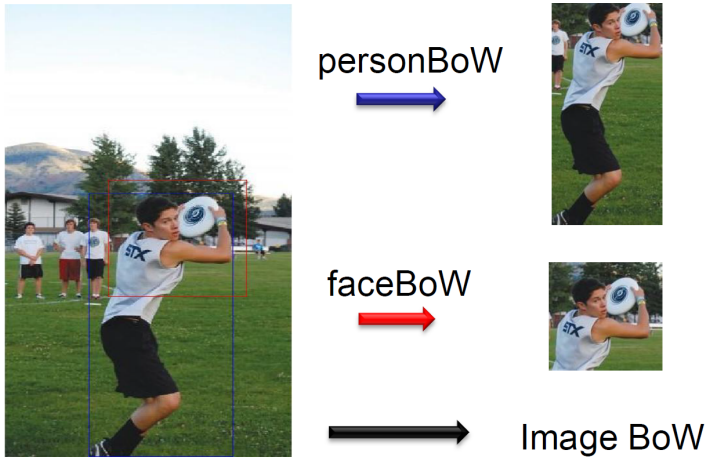
- Formation of bags from the still images.
  - b) Sample 100 windows from image
  - c) k-means over the appearance features and group into 10 clusters.
  - d) Form our bags



Additional Features

- Facial features
- Person appearance features
- Global image features





Comparison to state-of-the art

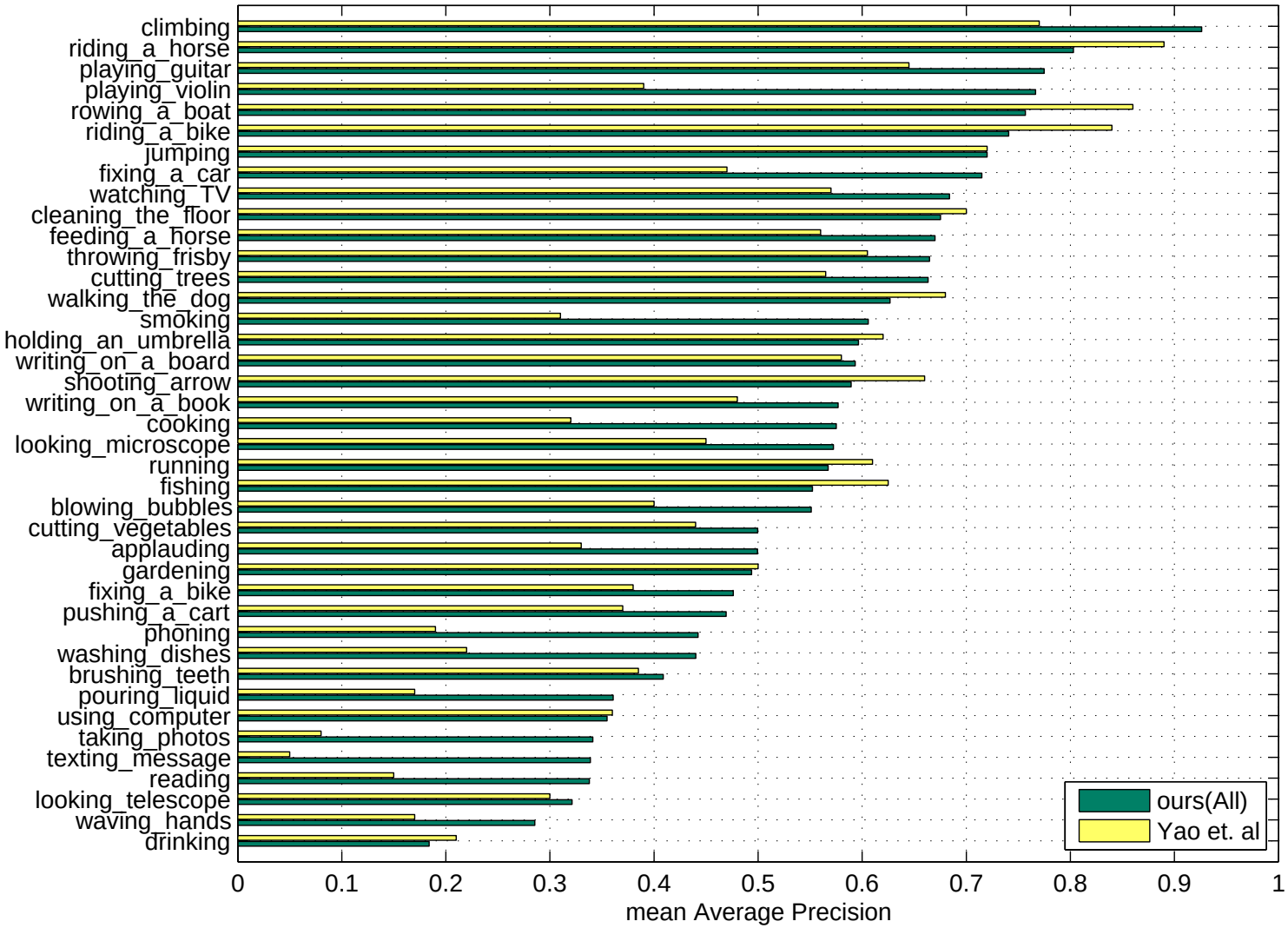
- Accuracy and mean average precision(mAP) of individual features and the combinations.
- Best performance ; MIL framework over the candidate object regions.

	accuracy	mAP
personHOG	24.75	19.35
personBoW	28.56	21.53
faceHOG	14.01	10.37
faceBoW	17.93	13.83
imgBoW	33.51	26.32
objectMIL	51.34	51.80
imgBoW+objectMIL	52.30	52.23
All(w/o objectMIL)	41.47	36.63
All	**55.93**	**55.55**
[Yao ICCV'11]	NA	45.7

[Yao ICCV'11] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in International Conference on Computer Vision (ICCV), 2011.

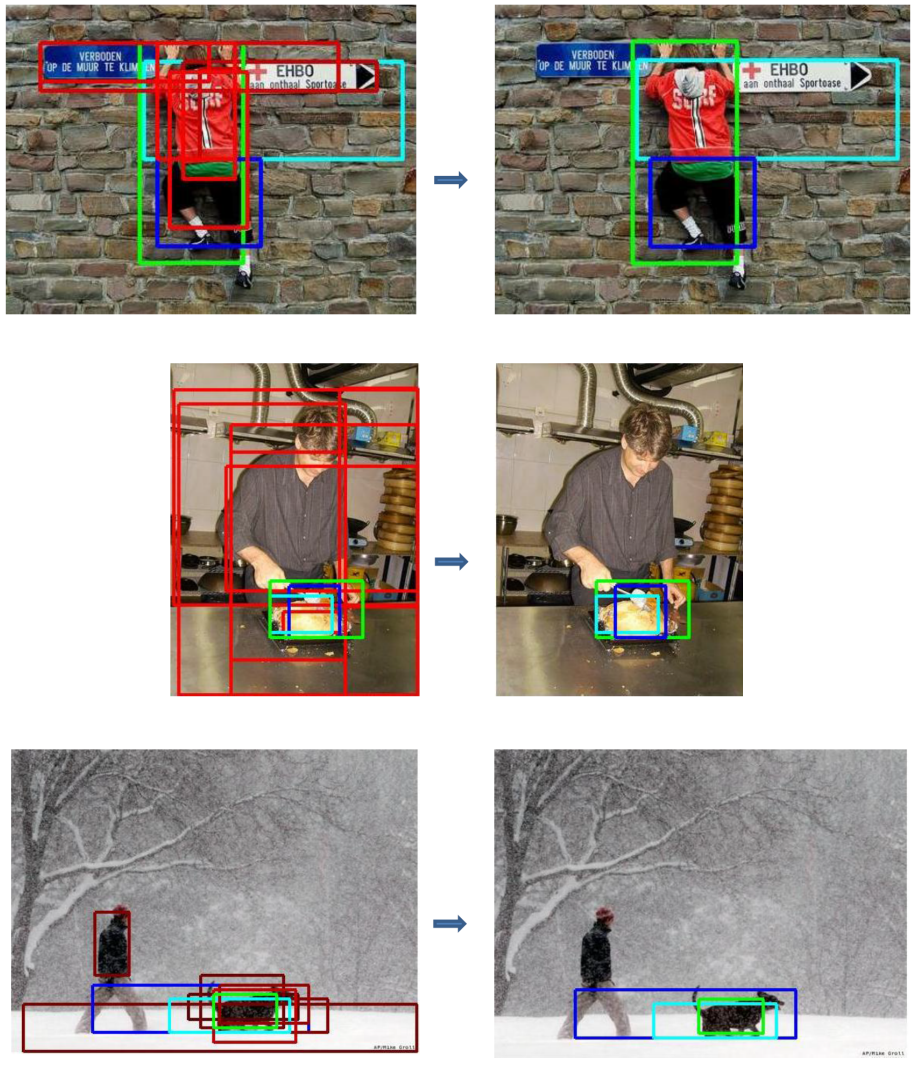
Comparison to state-of-the art

- Per action mAPs for each of the features. Overall, combining all the features' responses works the best



Most contributing concept instances

- Most contributing object/image regions discovered by the MIL

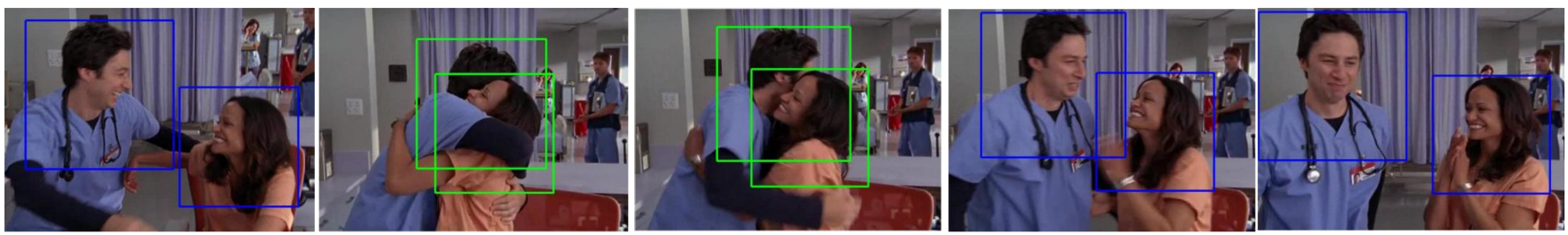


Recognizing human interactions using Multiple Instance Learning

- Problem Definition
- Related Work
- Our Approach
  - Modeling Person-Person Relationships
  - Image Representation
  - Bag Construction
  - MIL Classification and Spatial Embedding
  - Performance of Individual Features
  - Performance of Spatial Embedding
  - Comparison to state-of-the art

Problem Definition

- Recognizing human interactions from Real-world videos
- Real-world videos are weakly annotated;
  - A video has a class label
  - Interaction occurs in which frame?



Related Work

- Interactions in videos studied in a hierarchical manner, heavily depend on low level processes such as background subtraction, body parts etc. **[Datta 2002]**, **[Park 2006]**, **[Park 2009]**.
- Social interactions related faces. **[Fathi CVPR'12]**.
- Group interactions **[Lan CVPR'12]**.
- Interaction recognition in still images **[Yang CVPR'12]**. (personal space between individuls when they are interacting).

- **[Patron-Perez BMVC'10]** study recognizing interactions between two people real world video data, introduce a person-centered descriptor.
- Two-person interactions using multiple instance learning. **[Yun CVPR'12]**

Modeling Person-Person Relationships

- Orientation of faces, shape, motion, distance between indiviaduals.
- Form person-centered descriptors; body and face regions.
- Concatenation of two person region features

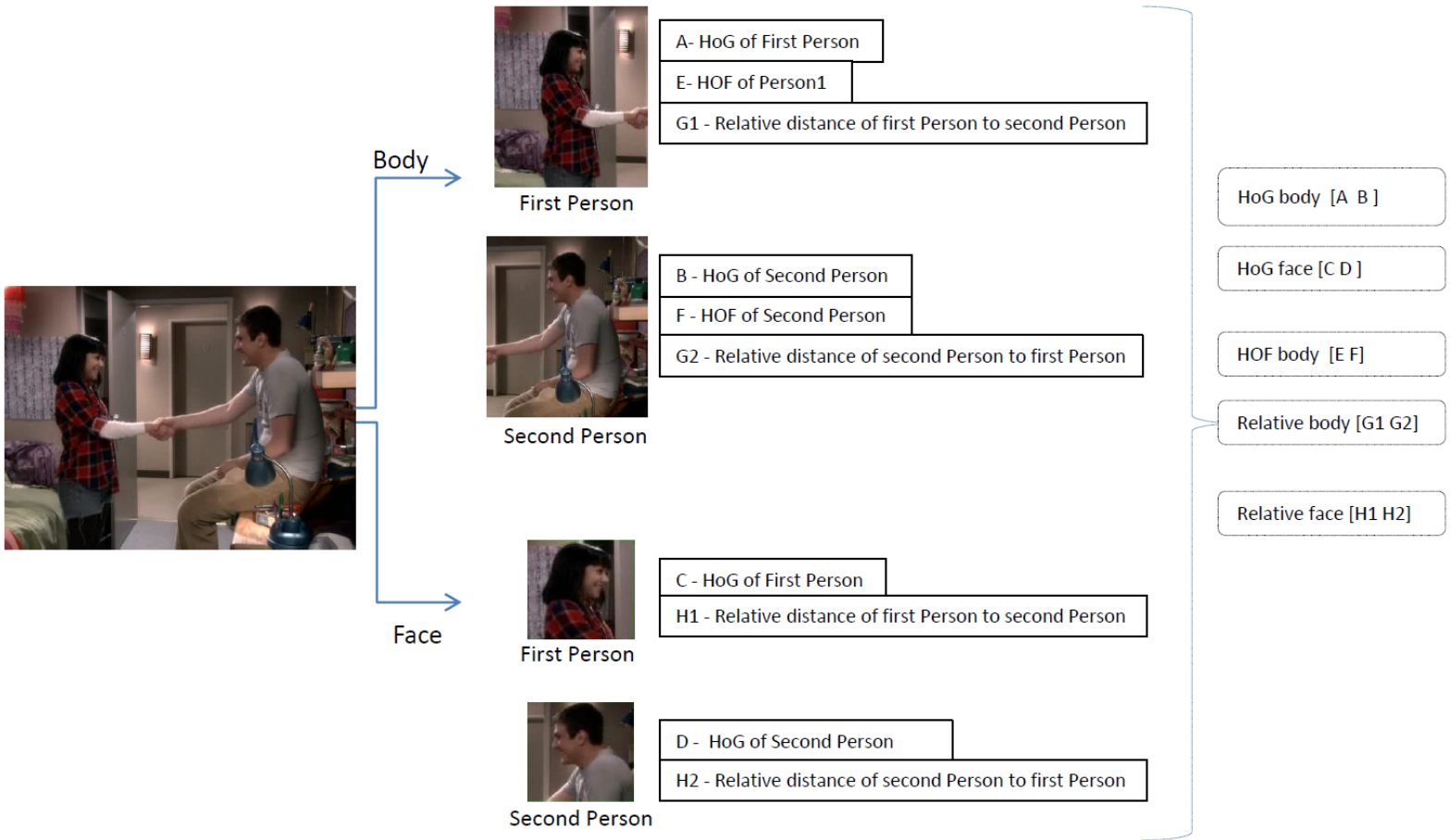
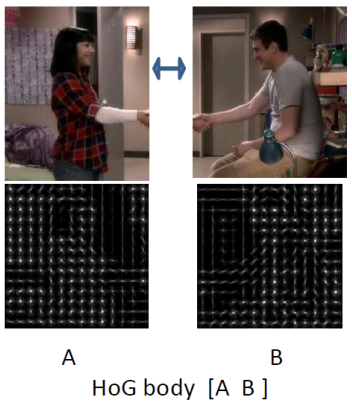
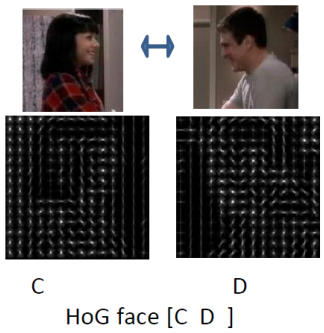


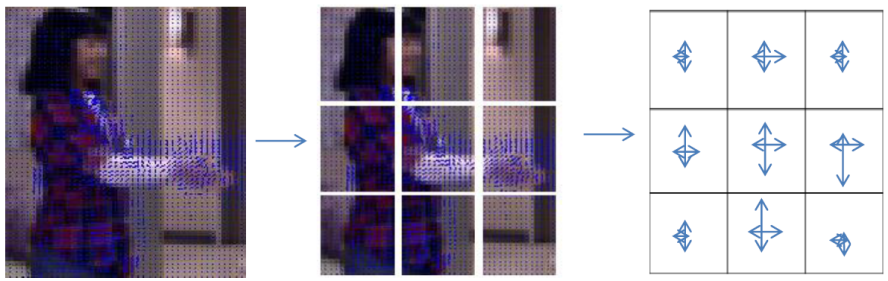
Image Representation

- Features from face and body regions of two people

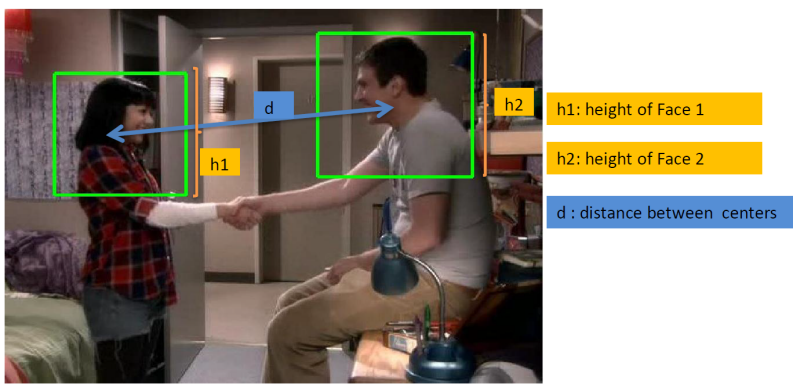




(h) Shape



(i) Motion



(j) Spatial relations between individuals

Bag Construction

- Video based Evaluation

Figure : Features extracted from first and second person region are concatenated. Then added to bag as an instance. Bag has two instances with red squares where interaction occurs.

Bag Construction

- Video based Evaluation

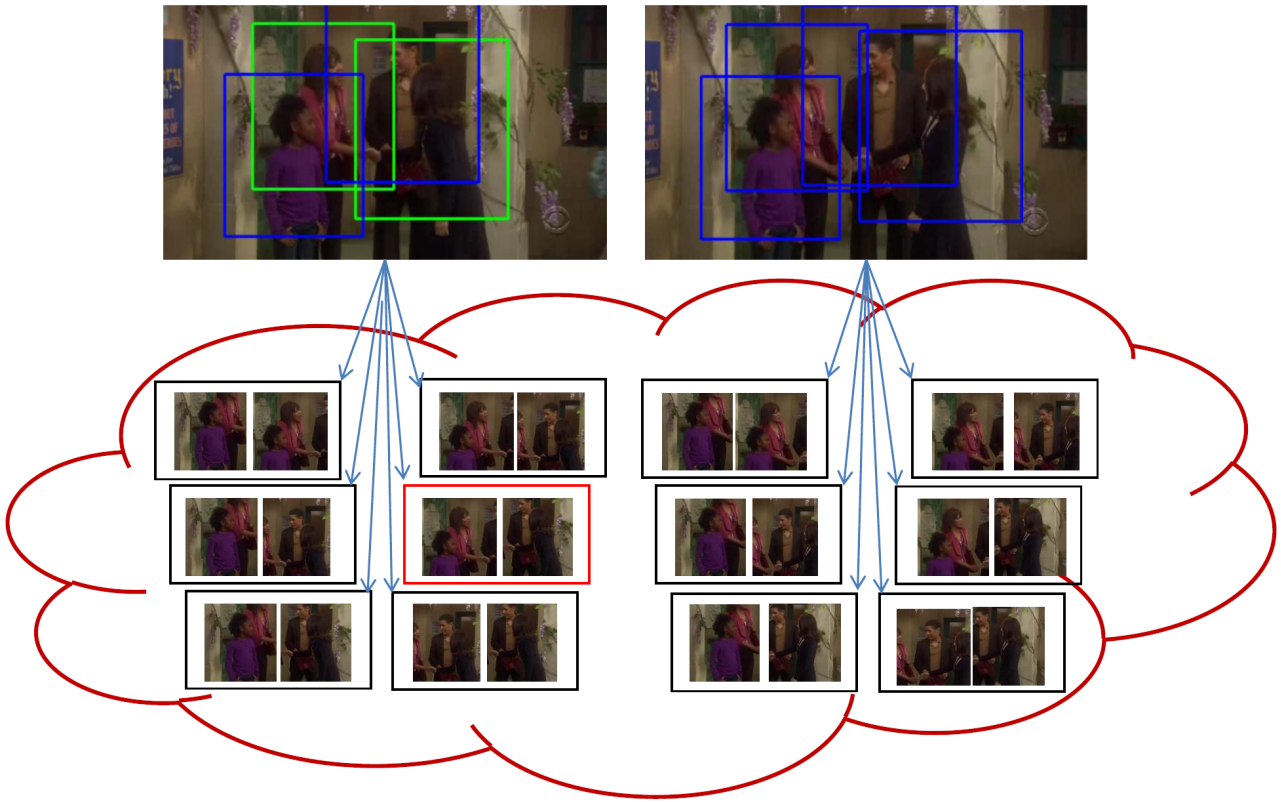


Figure : Frames include multiple people. There are 2 sample frames from an interaction video selected to construct a bag. Starting from leftmost person a match is done over other person regions stay in right-side. Bag has two instances with red squares where interaction occurs.

Bag Construction

- Frame based evaluation

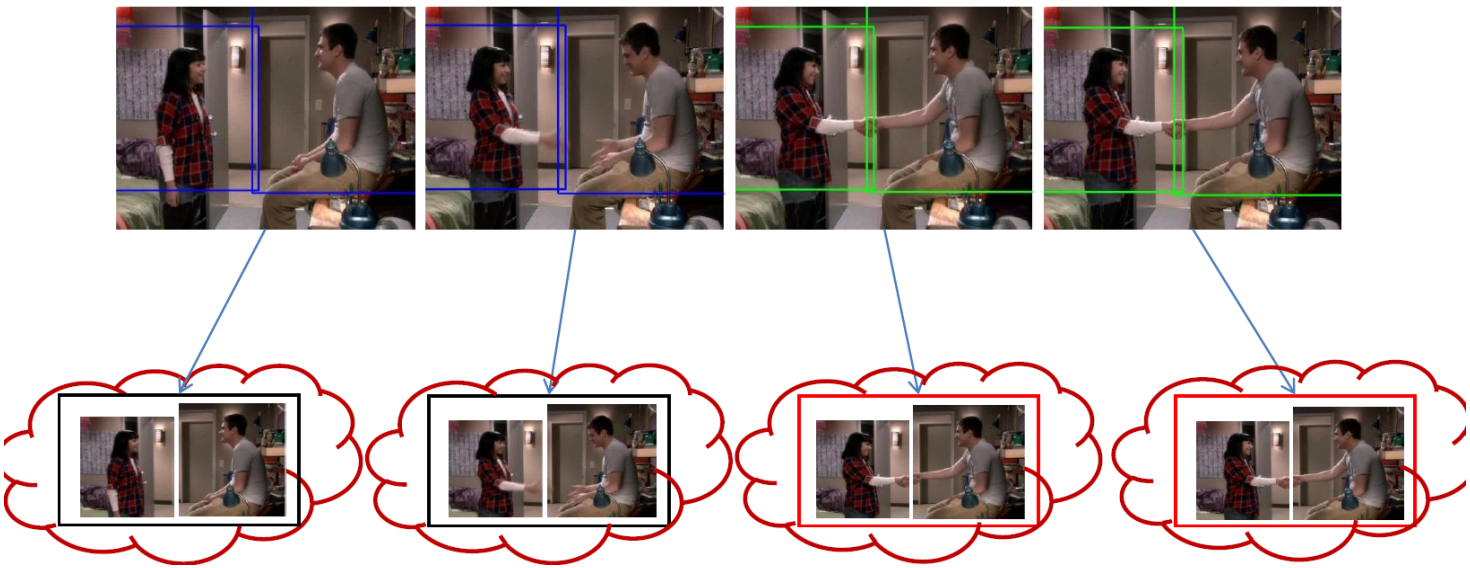


Figure : Features extracted from first and second person region are concatenated. For every match a bag is created. In two bags with red squares where interaction occurs. .

MIL Classification and Spatial Embedding

- We propose an extension to the spMILES **[1]**
- Spatial relations of person regions. Add two multiplicative spatial kernels to the feature-based similarity

$$s(c_l, \mathbf{B}_i) = \max_j \left(\phi_{feat}(x_{ij}, c_l) \phi_{spatial_x}(x_{ij}, c_l) \phi_{spatial_y}(x_{ij}, c_l) \right), \quad (2)$$

$$\phi_{feat}(x_{ij}, c_l) = \exp\left(-\frac{D(x_{ij}, c_l)}{\sigma} \right), \quad (3)$$

- where
  - $\phi_{feat}(x_{ij}, c_l)$ is the similarity between feature vectors.
  - $\phi_{spatial_x}(x_{ij}, c_l)$ is the spatial closeness between a concept instance c_l and a bag instance x_{ij} in x dimension
  - $\phi_{spatial_y}(x_{ij}, c_l)$ is the spatial closeness between a concept instance c_l and a bag instance x_{ij} in y dimension.

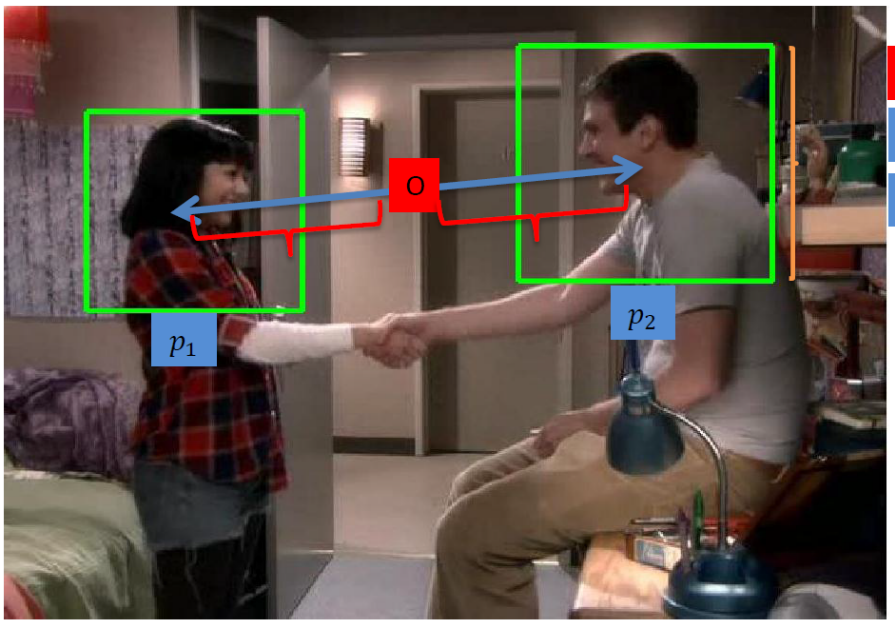
[1] N. Ikizler-Cinbis and S. Sclaroff, "Object recognition and localization via spatial instance embedding," in Pattern Recognition (ICPR), 2010

MIL Classification and Spatial Embedding

- Spatial relations of person regions. Add two multiplicative spatial kernels to the feature-based similarity

$$\phi_{spatial_x}(x_{ij}, c_l) = \exp\left(-\frac{|d_x(p_1, O) - d_x(p_1\,', O\,')|\,|d_x(p_2, O) - d_x(p_2\,', O\,')|}{\sigma_x}\right), \quad (4)$$

$$\phi_{spatial_y}(x_{ij}, c_l) = \exp\left(-\frac{|d_y(p_1, O) - d_y(p_1\,', O\,')|\,|d_y(p_2, O) - d_y(p_2\,', O\,')|}{\sigma_y}\right), \quad (5)$$



Performance of Individual Features - MILES

- Video based Evaluation on "TV Interactions" dataset**[1]** : Average Precision values of bag based evaluation method for 5 types of features individually. Negative video data is not included. (combination with equal weights and with weights provided by SVMs over training scores of each feature.)

method	hs	hf	h	k	mean
hog_body	63.08	52.06	62.24	68.64	64.32
hog_face	58.11	63.60	74.13	75.97	**66.15**
hof_body	59.90	63.99	49.64	49.25	59.47
relative_body	55.73	53.54	66.53	56.92	61.39
relative_face	54.40	54.01	67.73	61.37	62.50
equal weights	62.52	74.92	81.30	73.06	**72.95**
svm weights	65.46	73.48	76.28	75.61	72.71

[1] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High Five: Recognising human interactions in TV shows," in British Machine Vision Conference, 2010.

Performance of Spatial Embedding - spMILES

- Video based Evaluation on "TV Interactions" dataset**[1]** : Average Precision values of bag based evaluation method for three features with embedded spatial relations. Negative video data is not included. (combination with equal weights and with weights provided by SVMs over training scores of each feature.)

method	hs	hf	h	k	mean
hog_body_spatial	66.90	69.99	77.50	74.48	72.22
hog_face_spatial	63.57	66.47	87.06	84.18	**75.32**
hof_body_spatial	67.72	69.61	69.62	55.60	65.64
equal weights	70.25	70.29	84.18	80.25	**76.24**
svm weights	68.57	70.03	83.68	80.13	75.60

[1] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High Five: Recognising human interactions in TV shows," in British Machine Vision Conference, 2010.

Comparison to state-of-the art

- Comparison to state-of-the art on "TV Interactions" dataset over Average Precision values , V : video based , F: frame based, N : negative data, S : spatial embedding

method	hs	hf	h	k	mean
V	62.52	74.92	81.30	73.06	72.95
F	64.87	63.95	71.49	68.87	67.29
V + S	70.25	70.29	84.18	80.25	**76.24**
F + S	74.83	68.24	82.40	73.38	74.71
Patron-Perez el.al.**[1]**	57.83	51.08	71.16	76.54	64.15
V + N	49.65	57.63	82.39	77.35	**66.76**
Patron-Perez el.al.**[1]** + N	45.30	45.07	62.00	70.58	55.74

[1] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High Five: Recognising human interactions in TV shows," in British Machine Vision Conference, 2010.

Qualitative Results

- Video based Evaluation : Highest ranked true and false positives for `hog_body`, features. Ordering is done based one Average Precision values obtained from bag based evaluation. Negative video data is not included.

Conlusion

- We introduce novel solutions for three computer vision problems by successfully applying Multiple Instance Learning;
  - Image retrieval and re-ranking,
  - Recognizing actions from still images,
  - Recognition of human interactions in videos.

- Each of the problems are tested on benchmark datasets of the problems and compared with the state-of-the-art.
- The experimental results verify the advantages of the proposed MIL approaches to these vision problems

- We show if abstracting the visual problem to multi-instance representation deceration is done wisely The Multiple Instance Learning (MIL) paradigm arises to be very useful.

Thank You.

Thank You.

Any questions?