JOINT POSE ESTIMATION AND ACTION RECOGNITION IN IMAGE GRAPHS

Kumar Raja**, Ivan Laptev*†*, Patrick Pérez** *and Lionel Oisel**

* Technicolor Research and Innovation, Cesson-Sévigné, France
† INRIA - Willow Project, Laboratoire d´Informatique, École Normale Supérieure, France

ABSTRACT

Human analysis in images and video is a hard problem due to the large variation in human pose, clothing, camera view-points, lighting and other factors. While the explicit modeling of this variability is difficult, the huge amount of available person images motivates for the implicit, data-driven approach to human analysis. In this work we aim to explore this approach using the large amount of images spanning a subspace of human appearance. We model this subspace by connecting images into a graph and propagating information through such a graph using a discriminatively-trained graphical model. We particularly address the problems of human pose estimation and action recognition and demonstrate how image graphs help solving these problems jointly. We report results on still images with human actions from the KTH dataset.

Index Terms— Action Recognition in still images, Pose estimation, Graph optimization

1. INTRODUCTION

We address the problem of human action recognition and pose estimation in still images. While human action recognition has been mostly studied in video, actions provide valuable description for many static images, hence, automatically identifying actions in such images could greatly facilitate their interpretation and indexing.

Human action recognition is known to be a hard problem due to the large variability in human pose, clothing, viewpoints, lighting and other factors. Identifying actions in still images is particularly challenging due to the absence of motion information helping action recognition in video. Several works have addressed human analysis in still images by identifying body pose [1, 2, 3]. In particular, methods addressing human pose estimation and action recognition jointly have been recently proposed in [4, 5] motivated by the interdependency between the pose and the action. Such methods, formulated in terms of graphical models, are typically trained on manually annotated examples of person images and are then applied to individual images during testing.

The number of available annotated training images is usually limited due to the high costs associated with the manual annotation. At the same time, huge collections of images with no or noisy labels are now available online approximating the dense sampling of the visual world. Such collections have been successfully explored by recent work on object and scene recognition [6, 7] and in graphics [8].

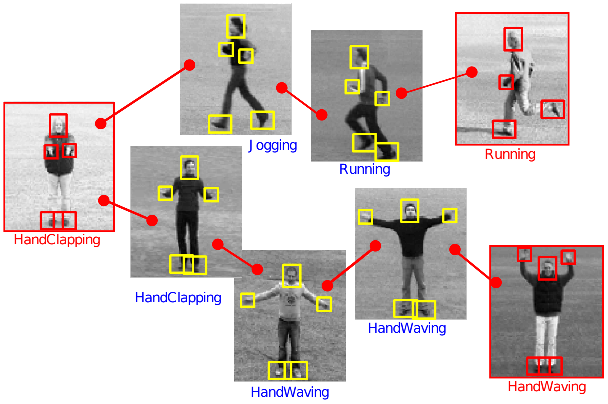


Fig. 1. Joint pose estimation and action recognition in the image graph. Training images (red frames) are manually annotated with the position of body parts and action labels. Part positions and action labels in test images (yellow frames) are resolved by optimizing the global graph energy.

In this paper we aim to push the above ideas further and to explore dense image sampling for human analysis. We assume a large number of images is available spanning the subspace of particular human actions. We assume only some of these images are annotated and use the remaining images to propagate information between each other. The underlying assumption behind our method is that images with small distance in the image space will often have similar semantics such as human pose and actions. We formalize this intuition in a graphical model by connecting similar images of people in a graph as illustrated in Fig. 1. We in particular, address the problems of human pose estimation and action recognition and demonstrate how the proposed image graphs enable to improve solutions for both of these tasks when solved jointly.

Related work. Action recognition in still images was addressed by Ikizler *et. al* [9] who used histogram of oriented rectangles as features and SVM classification. In [10] action images were collected from the web using text queries and an action model was built iteratively. Actions in consumer photographs were collected and recognized in [11] using Bag-of-

Words and LSVM classifiers.

Several other methods attempted action recognition by explicitly modeling the structure of the human body and its relation to manipulated objects. Graphical models have been used in [1, 2, 3] to model relations among body parts. More recently, [4, 5] extended this work towards the joint modeling of human poses and actions. We build on top of this work and extend it by leveraging the large number of unlabeled images. In this regard our work is related to the methods of object and scene recognition using large collections of unlabeled images [6, 7] and extends it to human analysis.

Overview. First, we describe our joint graphical model for human pose and actions in a single image. The graph energy is defined in terms of image-dependent and image-independent terms as described in Section 2. Next, Section 3 presents our contribution by extending the single-image graphical model to multiple images. This extension allows us to exploit unlabeled images while solving for the poses and actions in all images simultaneously. Experiments validating our approach are reported and discussed in Section 4. Finally, Section 5 concludes the paper.

2. JOINT MODEL FOR A SINGLE IMAGE

Motivated by the idea of pictorial structures [12] and following previous work [1, 2, 3, 4, 5], we model people using graphs. Our graphical model of a person (see Fig. 2(a)) contains six variable nodes encoding the positions of five body parts and the action label. We consider body parts $p \in \mathcal{P} = \{H, RH, LH, RF, LF\}$ corresponding to head, right-hand, left-hand, right-foot and left-foot, respectively, as well as K action classes A. The links between the nodes encode action-dependent constraints on the relative position of body parts and their appearance. Fig. 1 illustrates action labels and positions of the five body parts for some of our samples. Note that in some samples (depending on viewpoints) some parts may not be visible. So we include "occlusion" as one of the possible states of the part nodes. Using the part positions $x^p, p \in \mathcal{P}$, the pose vector is defined as $P = [x^p]_{p\in\mathcal{P}}$.

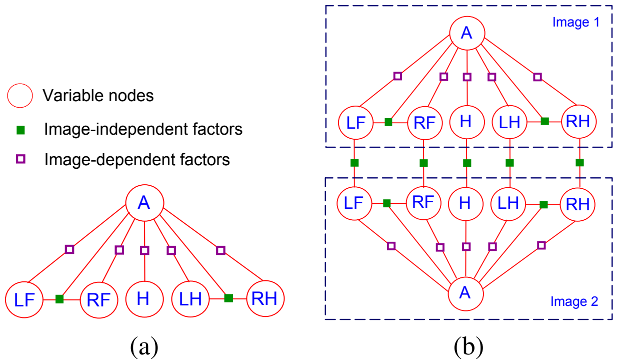
We define the energy of our pose-action graph in terms of image-dependent and image-independent potentials using pose P and action label A as variables:

$$E(A,P) = \phi(A;I) + \sum_{p\in\mathcal{P}} \phi_p(x^p, A; I) + \psi_h(x^{RH}, x^{LH}, A) + \psi_f(x^{RF}, x^{LF}, A) \quad (1)$$

where image-dependent potentials $\phi(.;I)$ encode appearance of parts and actions in the image I, and image-independent potentials $\psi(.,.,A)$ encode relations between the body parts for an action. We estimate the pose and action by maximizing the energy E, over action labels and part locations:

$$(A^*, P^*) = \arg\max_{(A,P)} E(A,P) \quad (2)$$

The arguments which maximize the above expression are found by the *max-sum* algorithm [13]. The advantage of the proposed joint action-pose model will be shown experimentally in Section 4.



(a) (b)

Fig. 2. Graphical models. (a) Pose-action graph for a single person. (b) Pose-action graph for two people in similar pose.

Discretization of Pose Space. Pose estimation involves the maximization of (2) over all possible poses. The concatenation of five part positions as the pose vector makes the pose space huge and the maximization intractable. In order to overcome this, we discretize and narrow down the pose space by identifying the most probable locations of the part. To this end, we take several frames from different action videos in which the parts positions are annotated relative to the person bounding box. For each part, we form a set by combining the list of locations and random perturbations of them. This set constitutes the set of part node states and it drastically reduces the cardinality of the pose space. Fig. 3(a) illustrates the locations corresponding to the discrete states of the "head" and "righthand" part nodes. The number of states in the head and righthand nodes is 364 and 1921, respectively.

Image-dependent potentials. There are six image-dependent terms in (1). To model action potential $\phi(A;I)$, we learn a binary static-image action classifier for each of the K action classes using histogram of gradients (**HoG**)-features and LSVM detector [14]. For a given image we obtain K action scores, we scale them linearly to the interval $[0\ 1]$ and use them as values of $\phi(A;I)$.

Similarly, to define part potentials $\phi_p(x^p, A; I), p \in \mathcal{P}$, we train action-dependent body part detectors. In sample action images with five annotated body parts, we obtain HoG-based descriptors as features to train RBF-SVM parts detector using [15]. For each part node state (part location), the RBF-SVM detector score is evaluated and the scores scaled between 0 to 1 to get the potentials values $\phi(P, A; I)$. We fix the potential values corresponding to occlusion states to a constant corresponding to the frequency of a part being occluded among the training samples for a given action.

Image-independent potentials. Image-independent terms $\psi_h(x^{RH}, x^{LH}, A)$ and $\psi_f(x^{RF}, x^{LF}, A)$ define relations between positions of hands and feet in the image. We model them in terms of the discrete states of the part nodes.

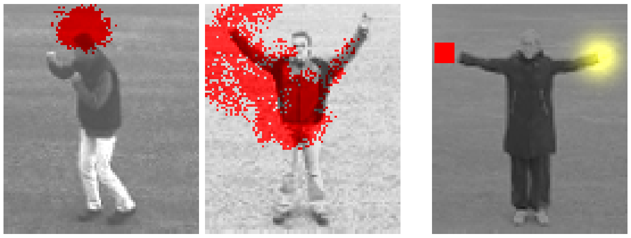


Fig. 3. (a): Discrete states (red) for positions of the head and the right-hand. (b): $\psi_h(x^{RH}, x^{LH}, A)$ showing possible left-hand locations (yellow) for handwaving action and a fixed right-hand position (red)

Similar to the discretization of the pose space above, for every action, we accumulate histograms of the joint locations of hands and feet in training images. For every instance of the joint part positions and action class, we update the corresponding histogram bin as well as the spatially-close bins using Gaussian weighting. Normalized histogram values are used to model $\psi_h(x^{RH}, x^{LH}, A)$ and $\psi_f(x^{RF}, x^{LF}, A)$ potentials. These terms can be interpreted as being proportional to the joint probability density of the right- and left-part location for a given action. Fig. 3(b) illustrates an example of $\psi_h(x^{RH}, x^{LH}, A)$ for the handwaving action and a fixed position of the right-hand.

3. IMAGE GRAPH

Here we propose to extend the graphical model for a single-person image in the previous section to multiple images. The rationale behind this extension is to take advantage of the large number of unlabeled person images and to propagate pose and action information among similar images. Such propagation should enable improved interpretation of samples when pose and action cannot be well-inferred from the training samples alone.

Given an image graph we introduce links between body parts of neighboring images. An example of a two-image pose-action graph is illustrated in Fig. 2(b). We connect all labeled (training) images and unlabeled (test) images into a graph and estimate the pose and the actions for unlabeled images simultaneously by maximizing the graph energy. We define the energy of the image graph as a function of pose and action in unlabeled images as

$$E^G(\{A\}_i, \{P\}_i) = \sum_i E_i(A_i, P_i) + \sum_{i\sim j} \sum_{p\in\mathcal{P}} d_p(x_i^p, x_j^p), \quad (3)$$

where $P_i = [x_i^p]_{p\in\mathcal{P}}$, E_i is the graph energy for a single image i defined in (1), and $d_p(x_i^p, x_j^p)$ are neighbor potentials between images i and j penalizing the placement of matching parts p far apart. We define $d_p(x_i^p, x_j^p)$ to be $\exp(-||x_i^p - x_j^p||^2/2\sigma^2)$, where σ is the spread parameter. Potentials $d_p(x_i^p, x_j^p)$ provide a strong prior on the position of parts in test images with direct connections to the training images. $d_p(x_i^p, x_j^p)$ also encourages similar test images to have similar placement of parts. Note that action nodes of neighbor images are indirectly connected through the body-part nodes. This enables propagation of action information through the graph.

Graph Construction. To construct an image graph, we connect every image (labeled and unlabeled) to its four nearest neighbors in terms of a distance measure. As the pose and the action labels are fixed for the training images, there is no flow of information among them. We therefore first connect every labeled image to its similar unlabeled images, then every unlabeled image is linked to its neighbors among labeled and unlabeled images. Our graph construction requires the notion of image similarity which can be defined in different ways. To validate our approach, in this work we find image similarity between a pair of images using the "true pose distance" which is defined as

$$D(i,j) = \sqrt{\sum_{k\in\mathcal{P}_v} ||P_k(i) - P_k(j)||^2 + nC_o} \quad (4)$$

where $P_k(i)$ and $P_k(j)$ are the annotated positions of corresponding body parts in images i and j respectively, $\mathcal{P}_v$ is the set of visible parts in both images and C_o is the fixed occlusion cost penalizing a part with mismatched occlusion label. In the future work we plan to substitute $D(i,j)$ by a measure based on image information only. Note that the part positions in the unlabeled images are predicted from graph optimization.

4. EXPERIMENTS AND RESULTS

We validate our framework by estimating human pose and recognizing actions in still frames extracted from KTH dataset. The dataset contains images of multiple people performing six classes of actions: boxing, handclapping, handwaving, jogging, running and walking. Our training and test sets are separated by person identities and contain 461 and 328 cropped person images respectively. To train and test human pose estimation, we have manually annotated bounding boxes of head, hands and feet in all our images. A few samples and corresponding annotations from our dataset are illustrated in Fig. 1.

To evaluate action recognition, we measure accuracy by the ratio of correctly classified images to the total number of test images. To evaluate pose estimation, we report Average Precision (AP) values computed for the five body parts (head, hands and feet). We assume a body part is correctly localized if the overlap between its predicted and its true bounding boxes is greater than 0.5.[1] The detection score of a part corresponds to the sum of the terms in the graph energy in which the part location is a variable, i.e. it is the sum of the factors surrounding the part node in the pose-action graph. We do not distinguish between the left/right hands and left/right feet

[1] We measure the overlap between two bounding boxes a and b as the intersection over union: $|a \cap b|/|a \cup b|$.

in the evaluation. Our framework allows for prediction of occluded parts. We consider a false detection if a part location is predicted for an occluded part, and vice-versa.

We evaluate human pose estimation and action recognition for different settings of the graphical model. To show the advantage of the joint approach to pose estimation and action recognition, we first consider an independent solution. For this purpose, in experiment E1 the action in an image is simply $A^* = \arg\max_A \phi(A; I)$. The location of a part $p, p \in \mathcal{P}$ is predicted as $x^{p*} = \arg\max_{x^p} \max_A \phi(x^p, A; I)$.

In the second experiment E2 we extend E1 and evaluate the advantage of action-dependent pose estimation by considering potentials $\phi_p(x^p, A; I)$ while still not modeling the relative position of parts, i.e. setting $\psi(., ., A) = 0$ in (1). In the experiment E3 we consider the solution provided by maximizing the full energy E in (1) of the single image graph. Finally, in experiment E4 we evaluate our extension of the single-image graph model to the multiple image graph. We report the solution obtained by maximizing the energy E^G in (3).

The results for all four experiments E1-E4 are reported in Tables 1 and 2 for action recognition and pose estimation respectively. By comparing results of E1 and E2, we observe the advantage of modeling the appearance of actions and parts jointly in E2. Additional modeling of relative positions of body parts in E3 demonstrates an improvement both for the action recognition and pose estimation compared to E1 and E2. Finally, the image graph proposed in this paper results in a clear improvement of action recognition and pose estimation in E4 compared to the single-image graph optimization in E3.

Experiment	Action Recognition(in %)
E1. Action classifier	78.35
E2. $\psi(., ., A) = 0$	81.09
E3. 1-image graph	82.62
E4. N-image Graph	**86.58**

Table 1. Accuracy of action recognition.

Experiment	Head	Hands	Feet	mAP
E1. Part Detectors	0.9645	0.3854	0.5357	0.6286
E2. $\psi(., ., A) = 0$	0.9604	0.4105	0.6790	0.6833
E3. 1-image Graph	0.9608	0.5206	0.9256	0.8023
E4. N-image Graph	**0.9892**	**0.8293**	**0.9745**	**0.9310**

Table 2. Average precision for part localization.

5. CONCLUSIONS AND FUTURE WORK

We have extended a joint model of human action and pose to image graphs in a semi-supervised framework for human analysis. In this extension, inference on pose and action is performed in unlabeled images of an image graph using the image connections. A crucial aspect in this framework is the construction of the image graph in which neighboring images must have similar pose. The immediate future work to improve upon the presented ideas includes: 1) use of an automatic mechanism to determine the image distance used to construct image graphs, 2) building of image graphs with specific topologies to leverage the connections better, and 3) exploring a more realistic dataset of human actions such as consumer photographs.

6. REFERENCES

[1] D. Ramanan, “Learning to parse images of articulated bodies,” in *NIPS*, 2006.

[2] A. Gupta, A. Kembhavi, and L.S. Davis, “Observing human-object interactions: Using spatial and functional compatibility for recognition,” *IEEE PAMI*, vol. 31, no. 10, pp. 1775–1789, 2009.

[3] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, “Pose search: retrieving people using their pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[4] Weilong Yang, Yang Wang, and G. Mori, “Recognizing human actions from still images with latent poses,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, une 2010, pp. 2030–2037.

[5] Bangpeng Yao and Li Fei-Fei, “Modeling mutual context of object and human pose in human-object interaction activities,” in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010.

[6] A. Torralba, R. Fergus, and W.T. Freeman, “80 million tiny images: a large dataset for non-parametric object and scene recognition,” *IEEE PAMI*, vol. 30, no. 11, pp. 1958–1970, 2008.

[7] J.H. Hays and A.A. Efros, “Im2gps: estimating geographic information from a single image,” in *CVPR*, 2008.

[8] J.H. Hays and A.A. Efros, “Scene completion using millions of photographs,” in *Proc. ACM Transactions on Graphics (SIGGRAPH)*, 2007.

[9] Nazli Ikizler, R. Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu, “Recognizing actions from still images,” in *International Conference on Pattern Recognition(ICPR)*, 2008.

[10] N. Ikizler-Cinbis, R.G. Cinbis, and S. Sclaroff, “Learning actions from the web,” in *IEEE 12th International Conference on Computer Vision*, October 2009, pp. 995–1002.

[11] V. Delaitre, I. Laptev, and J. Sivic, “Recognizing human actions in still images: A study of bag-of-features and part-based representations,” in *Proc. BMVC.*, 2010.

[12] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, “Pictorial structures for object recognition,” *International Journal of Computer Vision*, vol. 61, pp. 55–79, January 2005.

[13] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, October 2007.

[14] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008, pp. 1–8.

[15] T. Joachims, *Making large-Scale SVM Learning Practical*, MIT Press, 1999.