ACTION RECOGNITION USING INSTANCE-SPECIFIC AND CLASS-CONSISTENT CUES

Chin-An Lin[1,2]*, Yen-Yu Lin*[1]*, Hong-Yuan Mark Liao*[3]*, Shyh-Kang Jeng*[2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[3]Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

We aim to resolve the difficulties of action recognition arising from the large intra-class variations. These unfavorable variations make it infeasible to represent one action instance by other ones of the same action. We hence propose to extract both *instance-specific* and *class-consistent* features to facilitate action recognition. Specifically, the instance-specific features explore the *self-similarities* among frames of each video instance, while class-consistent features summarize *within-class similarities*. We introduce a generative formulation to combine the two diverse types of features. The experimental results demonstrate the effectiveness of our approach.

Index Terms— Action recognition, video understanding

1. INTRODUCTION

Human action recognition has received strong attention in the fields of computer vision and video processing. As one of the most important components for video understanding, action recognition is essential to a wide range of applications, such as video surveillance, anomaly detection, and human-computer interaction. Despite the great applicability, a fundamental difficulty hindering the advance of action recognition is *large intra-class variations* [1]. Such variations can result from both intrinsic and extrinsic factors, such as posture differences among persons, ambiguities from clutter background, different camera perspectives, or partial occlusions.

1.1. Related work

Designing more robust feature representations for action recognition has gained significant progress recently. *Global representations*, in which the region of an action is encoded as a whole, are widely used for their simplicity. Bobick and Davis [2] extract silhouettes of an action and construct two templates, *motion energy image (MEI)* and *motion history image (MHI)*, to describe motions of silhouettes. Blank et al. [3] consider actions as 3D space-time shapes, and characterize those shape volumes by saliency and oriented features. However, global representations are very sensitive to self-occlusions, view variations, and noise. *Local representations*, especially the *bag-of-words* approaches [4, 5, 6], are popular recently. Approaches of this class compile histograms of local quantized features. Nevertheless, the geometric structure of the local features in the spatio-temporal space is ignored in these approaches. It often causes performance degradation.

A vast amount of research effort has been made on modeling the correlations among local features. Matikainen et al. [7] specify correlations by a frequency lookup table of quantized geometrical displacements. Prabhaka et al. [8] estimate the causalities between visual words, and integrate them as parts of feature representations. Sun et al. [9] detect and track interest points to yield trajectories, and adopt Markov chaining to analyze the transitions of local features. Besides, approaches in [10, 11, 12] apply graphical models, such as *HMM*, *MRFs*, or *CRFs*, to handling the temporal variations of human actions. However, strong independence assumptions are required in these graphical models. Hence it is almost infeasible to deal with long-term temporal dependency.

1.2. Our approach

In this work, we aim to address the difficulties caused by large intra-class variations in action recognition. Due to the intra-class variations, an action instance cannot be well described by other instances of the same class. Hence we propose to represent actions by extracting two disjoint types of features, i.e., *instance-specific* and *class-consistent* features. While the former investigate characteristics that may vary from instance to instance, the latter explore properties that commonly shared by instances of the same action.

Specifically, we develop a general method to capture temporal dynamics of an action via referencing the *self-similarities* among frames of an action sequence. *Multivariate linear prediction (MLP)* is adopted to aggregate all the causalities of previous frames on the current ones. The resulting action dynamics are instance-specific in the sense that frame is approximated by other frames of the same video instance. On the other hand, we use *support vector machines (SVMs)* to discover the *class-consistent* features based on the bag-of-words model. Finally, a generative formulation is introduced to integrate the two complementary types of features, and leads to a performance improvement of action

recognition.

2. THE PROPOSED FRAMEWORK

In the section, we first give the adopted formulation that simultaneously takes the two feature sets into account for action recognition. Then its two components that extract and model the two sets of features are described respectively.

2.1. Our Formulation

Suppose we are given a set of data of one particular action c, $S_c = \{X_i \in \mathbb{R}^{T\times D}, y_i = c\}_{i=1}^N$, where X_i is a video with T frames and each frame $X_i(t)$ is described by a D-dimensional vector. In this work, a *generative* approach is adopted to learn the class-conditional densities $p(X|c)$, as well as the class prior $p(c)$ from S_c. We assume information embedded in X can be divided into temporal dynamics and static spatial information, and the two types of information are conditionally independent when the class of X is known, i.e.,

$$p(X, c) = p(X|c)p(c) = p(X'|c)p(\mathbf{x}_\mu, \mathbf{x}_\sigma|c)p(c), \quad (1)$$

where $\mathbf{x}_\mu \in \mathbb{R}^{1\times D}$ and $\mathbf{x}_\sigma \in \mathbb{R}^{1\times D}$ are the mean vector and the standard deviation vector of the rows of X respectively, and $X' \in \mathbb{R}^{T\times D}$ is obtained by normalizing X with respect to $(\mathbf{x}_\mu, \mathbf{x}_\sigma)$.

To learn $p(X'|c)$, we investigate self-similarities among successive frames from *relative* measurement X' to model *temporal dynamics* of an action. On the other hand for $p(\mathbf{x}_\mu, \mathbf{x}_\sigma|c)$, we explore commonly-shared properties among videos in S_c from *absolute* measurement $(\mathbf{x}_\mu, \mathbf{x}_\sigma)$ to capture *static spatial properties*. The features obtained in the former part are thought of as *instance-specific*, while the ones in the latter part are *class-consistent*. This point will be clarified later. Both types of features are jointly considered for action recognition in a principled way.

Specifically, we estimate temporal dynamics by multivariate linear prediction (MLP), and learn a static spatial model by support vector machines (SVMs) for each action c. For a testing instance Z, we predict its class by

$$c^* = \arg\max p(X'|\boldsymbol{\theta}_c)p(\mathbf{x}_\mu, \mathbf{x}_\sigma|\boldsymbol{\theta}_c)p(c), \quad (2)$$

where $\boldsymbol{\theta}_c$ is the set of parameters of the learned MLP and SVMs for action c.

2.2. Temporal dynamics via MLP

Temporal dynamics are described via *self-similarities* among frames, i.e., the current frame can be approximated by a linear combination of its previous ones. The recovered action dynamics are thought of as *instance-specific* in that each frame is described by only frames belonging to the same video instance. To this end, we develop *multivariate linear prediction*, which is a multi-dimensional generalization of linear prediction [13]. Given a multivariate point-process X' with zero mean and unit variance, it is formulated by

$$\begin{aligned} X'(t) &= \sum_{k=1}^{K} X'(t-k)A_k + \mathbf{e}_c(t) = \bar{X}_t A + \mathbf{e}_c(t) \\ &= \underbrace{[X'(t-1)\ \ldots\ X'(t-K)]}_{\bar{X}_t} \underbrace{\begin{bmatrix} A_1 \\ \vdots \\ A_K \end{bmatrix}}_{A} + \mathbf{e}_c(t), \quad (3) \end{aligned}$$

where $A_k \in \mathbb{R}^{D\times D}$ is the coefficient matrix and $\mathbf{e}_c(t) \in \mathbb{R}^{1\times D}$ is the reconstruction error. Variable A is typically optimized by minimizing the *expectation energy of reconstruction error*, i.e.,

$$A_c^* = \arg\min_A E\left[\sum_t ||\mathbf{e}_c(t)||^2 + \lambda \sum_{i,j} A^2(i,j)\right], \quad (4)$$

where $E[\cdot]$ is the expectation operator over videos of action c. Optimization problem (4) is convex with L_2 regularization, and has a globally optimal solution. We use cross-validation to determine λ, and obtain A_c^* by solving an equivalent linear problem. The details of this part is omitted due to the space restriction.

With A_c^* by multivariate linear prediction and applying the *probability* operator to both sides of (3), we have

$$p(X'(t)|X'(t-1), ..., X'(t-K), A_c^*) = p(\mathbf{e}_c(t)). \quad (5)$$

We see in (5) that *high-order* temporal dependency, which is hardly modeled by graphical models such as HMM and MRF, is described by the probability of reconstruction error. It follows that term $p(X'|c)$ in (1) can be rewritten as

$$\begin{aligned} p(X'|c) &= \prod_t p(X'(t)|X'(t-1), ..., X'(t-K), A_c^*) \quad (6) \\ &= \prod_t p(\mathbf{e}_c(t)). \quad (7) \end{aligned}$$

In our implementation, $p(\mathbf{e}_c(t))$ is defined by

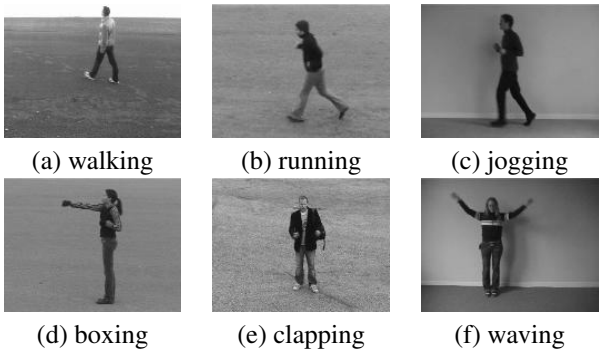
$$p(\mathbf{e}_c(t)) = \exp\left(-\gamma||\mathbf{e}_c(t)||^2\right), \quad (8)$$

where γ is a positive constant.

2.3. Static spatial model via SVMs

We learn $p(\mathbf{x}_\mu, \mathbf{x}_\sigma|c)$ in (1) via establishing the static spatial model of action c. Gaussian or Gaussian mixture models fulfill this goal, but with empirical comparison we find that better recognition performance of this part is obtained by adopting a discriminative classifier with Bayes' rule, i.e.,

$$p(\mathbf{x}_\mu, \mathbf{x}_\sigma|c) = \frac{p(c|\mathbf{x}_\mu, \mathbf{x}_\sigma)p(\mathbf{x}_\mu, \mathbf{x}_\sigma)}{p(c)} \propto p(c|\mathbf{x}_\mu, \mathbf{x}_\sigma). \quad (9)$$



(a) walking (b) running (c) jogging

(d) boxing (e) clapping (f) waving

Fig. 1. The KTH dataset. (a) $\sim$ (f) Six action classes.

From (2), we know $p(\mathbf{x}_\mu, \mathbf{x}_\sigma)$ is irrelevant to recognition. The class probability $p(c)$ is typically determined by one's prior knowledge and is set as a uniform distribution in our cases. As for $p(c|\mathbf{x}_\mu, \mathbf{x}_\sigma)$, we estimate it by learning an SVM classifier with probability outputs [14]. The yielded SVM classifier f_c is specified by decision boundary $\mathbf{w}_c$. The static model obtained in this way is considered as *class-consistent*, since f_c is developed to separate all videos of action c from the rest by exploring their commonly shared features. With temporal dynamics in (7) and static model in (9), the logarithm of (2) becomes

$$c^* = \arg\max_c -\gamma \sum_{t=1}^{T} ||\mathbf{e}_c(t)||^2 + \log f_c(\mathbf{z}_\mu, \mathbf{z}_\sigma; \mathbf{w}_c), \quad (10)$$

where γ weighs the importance tradeoff between action dynamics and static spatial model, and is set by cross validation.

3. IMPLEMENTATION DETAILS

The details of the adopted features for video description are given in the section. First a set of dense trajectories is generated from each video using the algorithm by Wang et al. [6] with parameter setting ($L = 14, n_z = 7$). Each trajectory is characterized by three different descriptors: 1) *motion boundary histogram (MBH)*, 2) *histogram of oriented gradient (HOG)*, and 3) *histogram of optical flow (HOF)*. Then a dictionary is constructed in the similar way as the one in [6] but with two differences: 1) We don't quantize a trajectory into one word but into multiple words according to its temporal separation. For example, a 14-frame long trajectory is temporally divided into 7 segments. Each segment (with 2 frames) is quantized into one word; 2) The dictionary size is set to 300 for the concern of computational efficiency. Finally, a video is represented by a point-process matrix [12] whose rows are temporally ordered histograms over words. In our experiments, time interval between two successive samplings is set to two frames in 25 fps video sequences.

Table 1. Recognition rates of various approaches.

Method	Recognition rate
Rodriguez et al. [15]	88.7%
Wang et al. [6]	94.2%
Chen and Aggarwal [12]	90.9%
Le et al. [16]	93.9%
Ours (instance-specific)	93.9%
Ours (class-consistent)	93.6%
Ours (combined)	**95.0%**

Table 2. The confusion matrix on the KTH dataset.

	walk	run	jog	box	clap	wave
walk	**100%**	0%	0%	0%	0%	0%
run	0%	**82.6%**	*17.4%*	0%	0%	0%
jog	0%	*9.0%*	**91.0%**	0%	0%	0%
box	0.7%	0%	0%	**99.3%**	0%	0%
clap	0%	0%	0%	2.8%	**97.2%**	0%
wave	0%	0%	0%	0%	0%	**100%**

4. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed approach on the *KTH* [5] action dataset, one of the benchmark datasets for action recognition. It consists of six action classes: walking, jogging, running, boxing, waving, and clapping. One example from each of the six classes is shown in Fig. 1. Each action is repeatedly performed several times by 25 subjects. These video sequences are recorded in four different scenarios: outdoors, indoors with scale variation, outdoors with different clothes. Totally 2391 samples are yielded. These samples display rich intra-class variations caused by the combination of subjects and scenarios. They nevertheless provide a good test bed to demonstrate the importance of using both instance-specific and class-consistent cues.

We follow the evaluation protocol in [5] for performance measure. For parameter setting, parameter λ in (4) is set for each action, so that the lowest energy of the reconstruction error over the validation set is obtained. The reconstruction errors of each action with respect to different values of λ are plotted in Fig. 2. As for γ in (8), it is determined via two-fold cross validation.

Our approach with three different settings, i.e., using instance-specific features, class-consistent features, and both of them, is compared with several published systems, including [6, 12, 15, 16]. The recognition rates are summarized in Table 1. We see that good performance is obtained by using only instance-specific features. It implies that the temporal dynamics can be well described by these instance-specific features, and are discriminant enough for action recognition. On the other hand, we observe that the two types of features

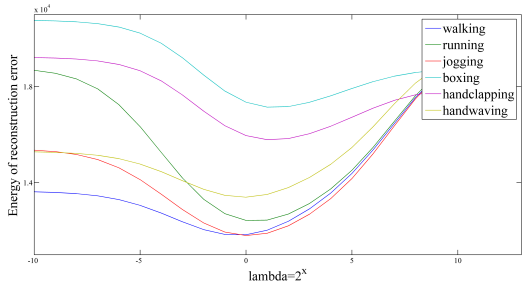


Fig. 2. The reconstruction errors vs. different values of λ for each action class.

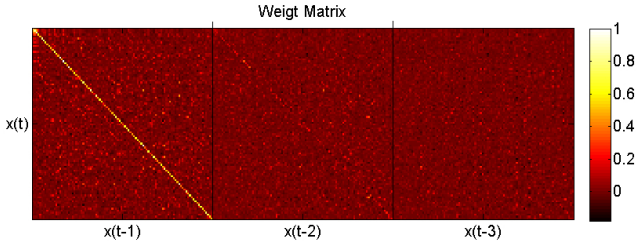


Fig. 3. The learned coefficient matrix $A_c^* = [A_1\ A_2\ A_3]$ in (4) for action walking.

seem to complement each other, and lead to a recognition rate of 95.0%. To our knowledge, this is the best recognition rate reported on KTH dataset.

Tabel 2 gives the confusion matrix. It is worth noting that relative to the state-of-the-art systems, such as [6, 12, 15, 16], the proposed approach has better ability to distinguish walking, jogging, and running. This reveals that it can effectively use the long-term dependency among successive frames to establish more reliable temporal dynamics. Finally, Fig. 3 visualizes the learned $A_c^* = [A_1\ A_2\ A_3]$ in (4) for action walking. It illustrates the causalities of the visual words.

5. CONCLUSIONS

We have presented a novel approach to action recognition by using both instance-specific and class-consistent cues. The instance-specific features are designed to describe action dynamics via MLP, while the class-consistent features conclude within-class similarities via SVMs. We introduce a generative formulation to elegantly combine the two diverse sets of visual features, and the proposed approach achieves the state-of-the-art recognition rate on KTH dataset. For future work, our approach will be more comprehensively evaluated under different settings and with various human actions.

Acknowledgments. The work is supported in part by grants NSC 100-2218-E-001-004 and 100-2221-E-001-013-MY3.

6. REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[2] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. Int'l Conf. Computer Vision*, 2005, pp. 1395–1402.

[4] I. Laptev and T. Linderberg, "Space-time interest points," in *Proc. Int'l Conf. Computer Vision*, 2003, pp. 432–439.

[5] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. Int'l Conf. Pattern Recognition*, 2004, pp. 32–36.

[6] H. Wang, A. Kläser, C. Schmid, and C.L. Liu, "Action recognition by dense trajectories," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3169–3176.

[7] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," in *Proc. European Conf. Computer Vision*, 2010, pp. 508–521.

[8] K. Prabhaka, S. Oh, P. Wang, G. D. Abowd, and J. M. Rehg, "Temporal causality for the analysis of visual events," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010, pp. 1967–1974.

[9] J. Sun, X. Wu, S. Yan, L.F. Cheong, T.S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 2004–2011.

[10] L. Wang and D. Suter, "Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007.

[11] J. Zhang and S. Gong, "Action categorization with modified hidden conditional random field," *Pattern Recognition*, vol. 43, no. 1, pp. 197–203, 2010.

[12] C.C. Chen and J.K. Aggarwal, "Modeling human activities as speech," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3425–3431.

[13] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 6, pp. 561–580, 1975.

[14] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[15] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach: A spatio-temporal maximum average correlation height filter for action recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.

[16] Q.V. Le, W.T. Zou, S.T. Teung, and A.Y. Ng, "Learning hierarchical invariant sptio-temporal features for action recognition with independent subspace analysis," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3161–3368.