Skeleton-augmented Human Action Understanding by Learning with Progressively Refined Data

Shih-En Wei* Nick C. Tang* Yen-Yu Lin* Ming-Fang Weng† Hong-Yuan Mark Liao*
* Academia Sinica, Taiwan
†Institute for Information Industry, Taiwan

ABSTRACT

With the aim at accurate action video retrieval, we firstly present an approach that can infer the implicit skeleton structure for a query action, an RGB video, and then propose to expand this query with the inferred skeleton for improving the performance of retrieval. It is inspired by the observation that skeleton structures can compactly and effectively represent human actions, and are helpful in bridging the semantic gap in action retrieval. The focal point is hence on action skeleton estimation in RGB videos. Specifically, an iterative training procedure is developed to select relevant training data for inferring the skeleton of an input action, since corrupt training data not only degrades performance but also complicates the learning process. Through the iterations, relevant training data are gradually revealed, while more accurate skeletons are inferred with the refined training set. The proposed approach is evaluated on ChaLearn 2013. Significant performance gains in action retrieval are achieved with the aid of the inferred skeletons.

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern Recognition

General Terms

Theory

Keywords

Pose estimation; Action retrieval

1. INTRODUCTION

Human action understanding has attracted great attention in the field of multimedia, because it is essential to widespread applications, such as human computer interaction and content-based video retrieval. A main goal of understanding an action is to acquire its semantic meanings. As actions are often considered as sequences of articulated poses in temporal order, the *skeleton structure*, which specifies the position configuration of key joints on the human body, gives a more compact and effective way for action representation. It is helpful in bridging the semantic gap in action retrieval. With the aim to enhance action retrieval, we propose to expand an action in RGB video with its estimated skeleton structure.

Although skeletons can be easily extracted with the aid of Microsoft Kinect, the short effect distance of Kinect, i.e., within four meters, often makes it not online available for capturing actions in many real-world applications, such as surveillance. We tackle the problem by offline collecting an auxiliary database, in which actions in RGB videos and their corresponding skeleton structures are available jointly. Such a database can be easily constructed by Kinect. It is compiled in advance, and covers actions of interest. In this work, the expanded skeleton for each query action will be estimated by referencing the auxiliary database.

However, inferring the implicit skeleton structure is very challenging owing to some unfavorable factors, such as self-occlusions and intra-class variationsBesides, the quality of the inferred skeletons is highly dependent on the degree of relevancy of training data, because irrelevant training data not only cause performance degradation but also complicate the training process. Therefore, we integrate relevant action selection into the procedure of skeleton model learning. It follows that more accurate skeletons are derived to facilitate action retrieval, the underlying goal of this work.

1.1 Related Work

Many powerful descriptors for capturing the characteristics of actions in RGB videos have been proposed. For example, the *space-time shapes* [4] extracted from silhouette images of human subjects and the 3D *Histogram-of-Gradient* (HOG) descriptor [10], which generalizes the spatial HOG features to the time domain. Although these descriptors achieved encouraging results in recognizing human actions, they are still not discriminative enough to deal with mutual or self-occlusion and inter-personal differences.

An action sequence can be viewed as a sequence of articulated poses along the temporal axis. Another trend of research is the development of *pose-based* approaches for better analysis of human actions. Methods of this category, e.g., [7, 8], carry out pose estimation to detect body parts in each frame, and recognize actions by analyzing the motions of body parts. These methods rely on the high-quality estimation of poses, which is a difficult task even if a large amount of training data is provided. For example, Wang et al. [8] proposed a method to select a best combination from

HuEvent'14 November 7, 2014, Orlando, Florida, USA

http://dx.doi.org/10.1145/2660505.2660512.

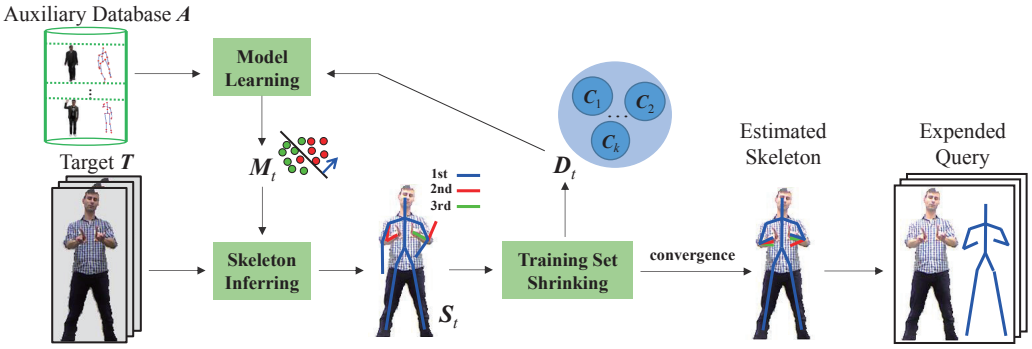

Figure 1: The Proposed framework for inferring the skeleton of an input action.

N-best estimations [11] by taking temporal correlation into account. However, this method still suffers from the noisy estimations and consecutive false estimations in a spatially nearby region. Motivated by the observation that a large portion of incorrect pose estimation results from irrelevant training data, our approach adopts an iterative procedure in which acquiring relevant actions and learning the skeleton model are done alternatingly.

1.2 Our Approach

Our approach is featured with the following two main contributions. First, to bridge the semantic gap in action retrieval, we augment an action depicted by RGB features with the inferred skeleton features. The two kinds of features catch diverse characteristics of actions, and the complementary information generally facilitates action retrieval. This step can be considered as a cross-modal generalization of *query expansion*: each query action in RGB features is associated with additional skeleton features. Second, we design an effective algorithm for estimating action skeletons. The algorithm achieves superior performance, because it iteratively retrieves a relevant subset of training data, and alleviates the unfavorable effects caused by a corrupt training dataset. After completing the estimation of the skeleton at each frame, the skeleton features of the video can be compiled by taking temporal dynamics into account. Besides, we will demonstrate that the resulting skeleton features are powerful and compact, and can lead to significant improvement for retrieving actions on the ChaLearn dataset.

2. SKELETON ESTIMATION WITH PROGRESSIVELY REFINED DATA

In this section, we describe our approach to skeleton estimation. Given an action T in an RGB video clip and auxiliary database A where actions in RGB videos as well as the corresponding skeletons are available, our goal is to infer the skeleton of T. The proposed iterative procedure is illustrated in Figure 1 for an overview. Each iteration consists of three stages, including *model training*, *skeleton inference*, and *training set shrinking*. At each iteration t, the objective of the first stage is to learn a model M_t for skeleton estimation with the training dataset D_t, which is initialized to A, and shrunk iteratively. The second stage is to infer the skeleton of T. We apply the learned model M_t to each frame of T, and retrieve multiple skeleton hypotheses of high scores. The set of the hypotheses S_t is then collected across frames. In the third stage, we partition D_t into k clusters. Only clusters that are relevant to S_t are kept and used to compile D_{t+1} for the next iteration. The three stages are performed iteratively until the training set is no longer shrunk. The three stages are respectively described below, followed by a brief justification. Note that we omit the subscript t for the sake of clearness.

2.1 On Learning the Skeleton Model

For skeleton model learning, we adopt the algorithm proposed in [11], where a pose (or a skeleton) $\mathcal{P}$ is specified by a *tree structure* (V, E) with K parts. Each part $v_i \in V$ is described by its label (e.g. head), location $l_i = (x_i, y_i)$, and type $t_i \in \{1, 2, ..., T_i\}$. Except for the root, each part v_i is connected to its parent part v_j, i.e., $e_{ij} \in E$. The model M consists of *appearance template* $\omega_i^{t_i}$ for each combination of parts and types, and *deformable parameters* $\omega_{ij}^{t_i}$ for modeling each part i and its parent part j in type t_i. These terms are jointly used for evaluating the compatibility of predicting pose $\mathcal{P}$ in image I through

$$S(I, \mathcal{P}) = \sum_{v_i \in V} \omega_i^{t_i} \phi(I, l_i) + \sum_{e_{ij} \in E} \omega_{ij}^{t_i} \psi(l_i - l_j), \qquad (1)$$

where $\phi(I, l_i)$ is the HOG features extracted at location l_i, and $\psi(l_i - l_j)$ is defined as $[dx\ dx^2\ dy\ dy^2]^\top$ with $dx = x_i - x_j$ and $dy = y_i - y_j$. For simplicity, we introduce the single-scale formulation here, but we adopt multi-scale one in the experiments. Refer to [11] for the details.

We slightly modify the model in [11] for improving the performance in our cases. The co-occurrence term is dropped. We individually train the model for each part v_i and type t_i, i.e., $(\omega_i^{t_i}, \omega_{ij}^{t_i})$. Besides, an extended *bootstrapping* procedure is applied to collect training data in deriving $(\omega_i^{t_i}, \omega_{ij}^{t_i})$. According to the data labels in D, we first gather a set of positive training data $f_p = \{(\phi(I, l_i), \psi(l_i - l_j))\}$. Then the set of the negative training data, f_n, is collected by applying bootstrapping to not only the negative videos but also the positive videos after excluding the regions of part v_i. With f_p and f_n, the model, $(\omega_i^{t_i}, \omega_{ij}^{t_i})$, for each part v_i and type t_i can be optimized based on the formulation of linear SVMs. *LIBLINEAR* [2] is used in our implementation.

2.2 On Inferring the Skeleton Structure

The task at the second stage is to infer the pose/skeleton in each frame of T. It can be efficiently done with the learned model M. The formulation in (1) allows us to use dynamic programming to determine the pose with the highest score.

Refer to [11] for the details. Because the estimated pose with the highest score may not be correct especially when the training data D are not highly relevant to T, we collect *multiple* pose hypotheses for each frame. Specifically, *N-best extensions* of dynamic programming and non-maximum suppression [6] are applied in this work. At this stage, we extract the top N estimations for each frame of T, and keep them in the set of skeleton hypotheses S.

2.3 On Shrinking the Training Dataset

The goal of this stage is to remove actions that are irrelevant to T from the current training set D. Specifically, we measure the dissimilarity between the skeletons in S and those in D, and shrink D by discarding skeletons less similar to those in S. To this end, we first describe the adopted representation for a skeleton $\mathcal{P}$. For each part v_i of $\mathcal{P}$, we normalize its position relative to the skeleton center by $l'_i = (l_i - l_s)/u$, where l_s is the center of the skeleton, and u is the length from the head to the spine and serves as the unit length. Without loss of generality, we assume that the root part is v_1, and compute the normalized position of part v_i relative to its parent part, i.e., $l''_i = l'_i - l'_j, \forall i \in \{2, 3, ..., K\}$. Then, the adopted skeleton features are defined as $\mathbf{s} = (l''_2, l''_3, ..., l''_K)$.

Note that since D (or A in the first iteration) is constructed from numerous consecutive video frames, the elements in D are quite redundant. While working on the whole set D is less efficient, we select farthest points [3] as training data to avoid selection bias caused by imbalance of data density across the feature space. To shrink D according to the relevancy to T, D is partitioned into k clusters with k-center clustering algorithm [3], i.e., $D = \bigcup_{n=1}^{k} C_n$ and $C_n \bigcap C_{n'} = \emptyset, \forall n \neq n'$. Then we shrink D by removing clusters that are less similar to the estimated skeletons of T. That is,

$$D \leftarrow \bigcup_{\mathbf{s} \in S} \arg\min_{C_n} d(\mathbf{s}, \mathbf{c}_n), \tag{2}$$

where $d(\mathbf{s}, \mathbf{c}_n)$ is the distance between the estimated skeleton $\mathbf{s}$ and the cluster head of C_n. The shrunk D will serve as the new training set at the next iteration.

The three stages are performed alternatively and iteratively until convergence. Namely, no elements need to be removed from D. At that time, we grab the most plausible skeletons for each frame of the target action T, and augment T with the grabbed skeletons for action retrieval. Through the iterations, we will show in the experiments that more accurate skeletons of T are inferred with the progressive refined training sets. We will also demonstrate the proposed cross-modal query expansion is helpful in boosting the retrieval performance.

3. FEATURE REPRESENTATIONS

The features representations adopted to characterize actions in RGB videos and skeleton structures for action retrieval are described as follows.

RGB Videos: For each action (query) T, we use or estimate its foreground mask to segment out the foreground region so that the space-time volume (STV) features from only foregrounds can be computed. The *3D-HOG (histogram of oriented gradients)* descriptor [10] is applied to describe the STV. In more detail, we use 16×16×16 pixel blocks, each of which is divided into 2×2×2 cells. Five hundred prototypes are derived to build up the embedding space.

Skeleton Structures: After grabbing the most plausible skeleton for each frame of query T, we implement *Fourier temporal pyramid* [9] to represent skeleton structures along the temporal domain. In this way, the temporal dynamics and coherence are taken into account. Specifically, the short time Fourier transform is applied to each skeleton segments in a 3-level pyramid. The feature representation is the concatenation of the Fourier coefficients from all the segments.

Table 1: The quality of the inferred skeletons, in terms of PCP_N and $NDCG_N$, along the iterations.

iter.	PCP_1	PCP_5	$NDCG_5$	PCP_{15}	$NDCG_{15}$
1	83.3%	76.1%	0.931	72.0%	0.904
2	88.1%	79.8%	0.947	76.0%	0.918
3	89.3%	80.7%	0.949	76.8%	0.922
4	89.7%	81.2%	0.950	77.1%	0.924

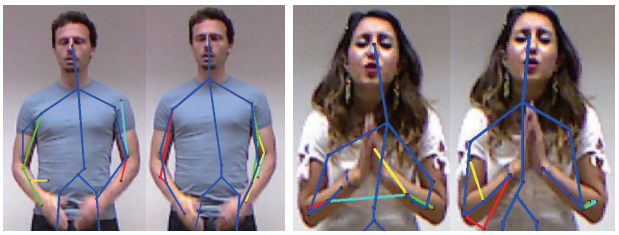

Figure 2: Two examples of the inferred skeletons at iterations 1 and 4. The top $N(= 5)$ estimations are specified with distinct colors, which are blue, green, red, yellow, and cyan in order.

4. EXPERIMENTAL RESULTS

In this section, our approach is evaluated on ChaLearn 2013 dataset [1]. We measure the quality of the inferred skeletons, and check whether expanding the query actions with the inferred skeletons improves retrieval.

4.1 The ChaLearn 2013 Benchmark

The ChaLearn dataset for gesture recognition contains 7,860 action video clips of 20 action types performed by 27 actors. The actions collected in this dataset include twenty kinds of Italian sign gestures. This dataset was recorded by using Kinect. The ground truth of the skeletons as well as foreground mask in each frame are available. To evaluated the performance of the proposed approach, we first select 800 action samples from the benchmark to yield the auxiliary dataset A, which consists of 20 actions performed by 10 actors, and each action was performed by four times. Except for those in auxiliary database A, each of the action videos in turn is treated as the query action, while the others form the database to be retrieved. In this setting, the performances of individual queries are measured, and average performance is reported.

4.2 Performance of Skeleton Estimation

We assess the quality of the inferred skeleton for each query. Total 20 joints are captured by Kinect to model skeletons. By combining the two closing endpoints on hands and foot, we use 16 joints (parts) in this work. Table 1 reports the accuracies of our approach along the iterative procedure of relevant data selection, when the top N skeleton

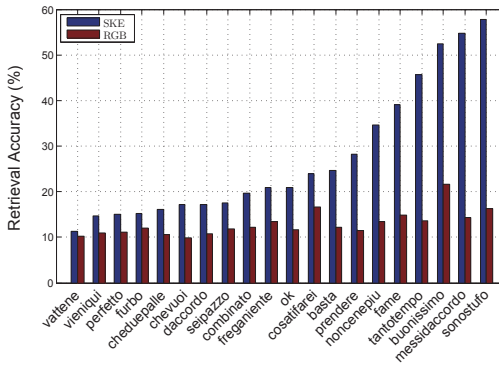

Figure 3: The retrieval accuracies on each of twenty classes of ChaLearn 2013 dataset.

estimations are considered with $N = 1, 5$, and 15, respectively. The accuracies are measured in terms of the metric *probability of correct part* (PCP) [11] and the *normalized discounted cumulative gain* measure (NDCG). PCP evaluates the quality of the estimated skeletons by comparing them with the ground truth. As can be found in Table 1, the average PCP converges in a few iterations. The gradually increasing PCP along the iterations manifest that our approach can effectively select appropriate training data, and infer high-quality skeletons. On the other hand, the progressively improved accuracies in NDCG point out that more accurate skeletons are assigned higher ranks after estimation. Figure 2 shows two examples of iteratively improved estimations of skeletons. It can be observed that all the top $N(= 5)$ estimations at iteration 4 are better than those at iteration 1 in both the two examples.

4.3 Performance of Action Video Retrieval

The effectiveness of the proposed cross-modal query expansion for action retrieval is assessed. The primary goal of the experiments is to check whether expanding the query action in RGB features with the estimated skeletons improves the retrieval performance. We follow [5] in computing the retrieval accuracy for each query, i.e., the precision when the top M $(= \lceil N/5 \rceil)$ results are returned, where N is the number of database entries whose classes are the same as that of the query. The retrieval accuracies on each of twenty classes of ChaLearn 2013 dataset are shown in Figure 4, where three approaches, i.e., `RGB`, `SKE`, are considered. In the first approach `RGB`, the similarity between the query action and entries in the retrieved database is measured by the features extracted from the RGB videos. As mentioned in Section 3, 3D HOG descriptor is adopted. In approach `SKE`, the similarity is measured based on their skeletons. It can be observed in Figure 4 that `SKE` significantly outperforms `RGB`. The performance gains of using the inferred skeleton features are remarkable in all the twenty classes especially in *combinato*, 41.9%(58.5%-16.6%), *cheduepalle*, 40.5%(55.3%-14.8%), and *fame*, 31.9%(45.9%-14.0%). To gain into the quantitative results, Figure 4 gives an example in which the query and the top four returns by `RGB` are shown. As can be seen that the skeleton features can alleviate the problems caused by the semantic gap between high-level concepts (actions) and low-level RGB features.

5. CONCLUSIONS

We have presented an effective approach that integrates

Figure 4: The query and the top 4 returns when only the query in RGB features is considered (upper) and when the query is augmented with skeletons (lower).

relevant data selection into the process of model learning, and can infer the skeleton structure of a query action accurately. The skeleton structure captures diverse information, and can be associated with the query action. The promising experimental results on ChaLearn 2013 benchamrk demonstrate that the proposed cross-modal query expansion leads to remarkable performance gains in action retrieval. Although skeletons of actions are easily extracted with the aid of Microsoft Kinect, some hardware limitations, such as the short effective distance and sensitive to infrared, make Kinect not online available in many practical applications. Our approach instead provides an alternative way of using Kinect even when it is not online accessible.

6. REFERENCES

[1] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, and H. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Int'l Conf. Multimodal Interaction*, 2013.

[2] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. Machine Learning Research*, 2008.

[3] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 1985.

[4] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Int'l Conf. Computer Vision*, 2005.

[5] S. Jones, L. Shao, J.-G. Zhang, and Y. Liu. Relevance feedback for real-world human action retrieval. *Pattern Recognition Letter*, 2012.

[6] D. Park and D. Ramanan. N-best maximal decoders for part models. In *Int'l Conf. Computer Vision*, 2011.

[7] K. Raja, I. Laptev, P. Perez, and L. Oisel. Joint pose estimation and action recognition in image graphs. In *Int'l Conf. Image Processing*, 2011.

[8] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *Computer Vision and Pattern Recognition*, 2013.

[9] J. Wang, Z.-C. Liu, Y. Wu, and J.-S. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition*, 2012.

[10] D. Weinland, M. Özuysal, and P. Fua. Making action recognition robust to occlusions and viewpoint changes. In *Euro. Conf. Computer Vision*, 2010.

[11] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition*, 2011.