Proceedings of IEEE Workshop on Applications of Computer Vision (WACV), Kona, Hawaii, January 2011

One Video is Sufficient? Human Activity Recognition Using Active Video Composition

M. S. Ryoo and Wonpil Yu
Electronics and Telecommunications Research Institute, Daejeon, Korea
{mryoo, ywp}@etri.re.kr

Abstract

In this paper, we present a novel human activity recognition approach that only requires a single video example per activity. We introduce the paradigm of active video composition*, which enables one-example recognition of complex activities. The idea is to automatically create a large number of semi-artificial training videos called* composed videos *by manipulating an original human activity video. A methodology to automatically compose activity videos having different backgrounds, translations, scales, actors, and movement structures is described in this paper. Furthermore, an active learning algorithm to model the temporal structure of the human activity has been designed, preventing the generation of composed training videos violating the structural constraints of the activity. The intention is to generate composed videos having correct organizations, and take advantage of them for the training of the recognition system. In contrast to previous passive recognition systems relying only on given training videos, our methodology actively composes necessary training videos that the system is expected to observe in its environment. Experimental results illustrate that a single fully labeled video per activity is sufficient for our methodology to reliably recognize human activities by utilizing composed training videos.*

1. Introduction

Human activity recognition is an active research area with a variety of applications. In the past few years, computer vision researchers have intensively explored recognition of actions, taking advantage of large-scale public datasets (e.g. the KTH dataset containing 2391 videos [16]). The paradigm of extracting features from a large set of training videos and statistically learning actions from them has been dominant, and many discriminative and generative approaches have been proposed [5, 10, 21]. Researchers have further extended this paradigm to analyze action videos with more complex backgrounds and camera movements, such as movie clips [9]. Complex human activities involving multiple persons (e.g. human-human fighting) have been recognized successfully as well [13], illustrating its applicability to surveillance systems. The assumption is that abundant training videos are provided to learn properties of human activities statistically.

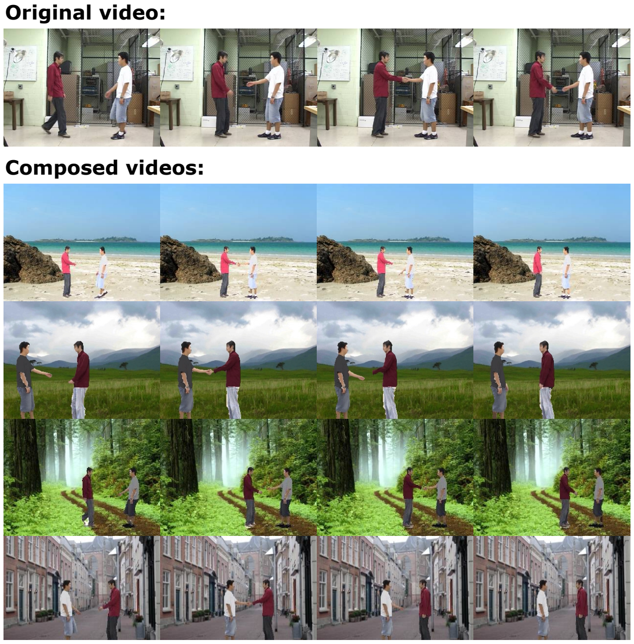

Figure 1. Example composed videos of 'shaking hands' with different backgrounds, locations, scales, actor clothings, and movement variations. The videos have been automatically generated by manipulating the original video. Note that the composed videos are showing various possible temporal structures of the shaking interaction : e.g. the person 1 is stretching/withdrawing his arm first, the person 2 is doing it first, they are doing it simultaneously, the person 1 is doing it more rapidly, and so on.

However, in many real-world environments (e.g. surveillance), we seldom have enough training videos to learn complex human activities. Activities have multiple variations, and we only have one or two exemplary training videos available per activity. For example, an abnormal activity of 'stealing' does not occur frequently, making its videos rare. This implies that collecting a large set of realistic training videos, which are required for the most of existing recognition approaches to capture activities' statistics, is difficult. A recognition system must possess an ability to process high-level activities from videos with various settings (e.g. illuminations and backgrounds) even when a small number of videos are given.

In this paper, we propose a methodology to learn and recognize human activities from an environment where only a single video per activity is provided. We introduce a new activity recognition paradigm that takes advantage of an active video composer, which automatically generates a large number of semi-artificial videos adaptive to its environment. Instead of solely relying on a single video to train the system, the idea is to compose various activity videos by manipulating the original video. Composed videos with actors having different clothing colors, locations, scales, and motion variations are generated while making their background to be identical to the testing environment. An active learning algorithm has been designed to ensure that the generated videos satisfy the temporal constraints of the activity. Figure 1 shows example composed videos.

The intuition behind our approach is to automatically generate diverse training videos that the system is expected to observe from the given scene. We represent an activity video as a composition of its scene background, actor locations, and the activity's sub-events, and present a methodology to generate composed videos given any arbitrary representation. That is, our video composer not only changes static components like the video's background from the original (i.e. real) video, but also modifies each sub-event's occurring time and speed to generate new videos with different structures. For example, 'shaking hands' videos are composed while making the 'stretching' of one person slow/fast, changing the temporal order of two persons' 'stretching's, and so on. Multiple activity videos with different temporal organizations are composed while considering other factors effecting video appearances such as clothings and scales.

An active learning algorithm has been designed to distinguish valid activity structure representations from invalid representations. The purpose is to make the video composer learn to discard abnormal video constructions and to provide only the videos with the correct activity. For example, in the interaction of a person 'pushing' another, the sub-event of one person being 'pushed away' must occur right after the other 'stretching an arm'. Otherwise (e.g. a person moved away 3 seconds *before* the other stretched his/her arm), the video violates the constraints of the activity, and becomes an abnormal video or a video with a different activity. Our algorithm learns to identify such abnormal structures, enabling the recognition system to take advantage of composed training videos with valid representations only.

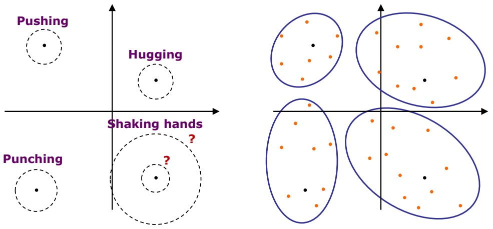

Figure 2. The left figure shows example subspaces of human activities, estimated using a single example per class. Black dots correspond to real training examples (i.e. videos), and dotted lines describe subspace boundaries. Conceptually, any point inside each region corresponds to a particular activity video. In the left figure, a subspace boundary of an activity is unclear (e.g. shaking hands), since only one example is provided. Our goal is to automatically generate the orange dots shown in the right figure (i.e. composed videos), so that the system is able to take advantage of them for more reliable learning of human activities.

The advantage of our video composition paradigm is that we are able to construct a large training set with multiple videos having different activity structures, even with a single video example. Once the learning is done, our video composer generates composed activity videos fully automatically. Only the videos satisfying the structural constraints of the activity are composed, and they are used for the training of the recognition system. Figure 2 illustrates the general idea behind our approach. Up to our knowledge, our paper is the first activity recognition paper to take advantage of composed videos. We also believe that our paper is one of the first papers to verify the benefits of active learning algorithms for the human activity recognition.

2. Related Works

Activity recognition. Human activity recognition has widely been studied by computer vision researchers since early 1990s [20]. Action recognition methodologies utilizing 3-D spatio-temporal (XYT) local features extracted from videos [5, 8] have particularly been popular in the past few years. Generative approaches constructing statistical models of actions [10, 21] as well as discriminative approaches searching for decision boundaries [5, 9, 13] have been developed, using the 3-D XYT features. Several public datasets have been established, mostly focusing on actions of single persons [16, 2]. In addition, high-level activity recognition approaches to analyze suspicious and abnormal activities have gained a large amount of interest [13, 12]. Many of these approaches assume that the representation of the human activities has been encoded by human experts, or require a large amount of training videos similar to the simple human action recognition cases.

Recently, there also has been an attempt to recognize human actions from a single example [17]. Their approach focuses on extracting specific feature patterns from a single video example, constructing one template per action. However, their system was limited in its ability to process complex human activities having motion variations: Complex humans activities with multiple actors often have several structural variations, making the extraction of the representative pattern from a single video infeasible.

Synthetic data. Even though the paradigm of using composed videos for the activity recognition has not been explored in depth previously, the idea of using fully synthesized data has been employed in other fields. Data mining researchers [4] have developed the idea of 'oversampling' data from a modeled probability distribution to increase the number of training samples from a minority class (which often is more important). There also has been an attempt to use sketched artificial images for training object recognition systems [18]. Furthermore, [11] constructed an artificial training set for a human interaction recognition, by extracting features directly from movements of synthetic agents in a 2-D plane. Even though the agent movements have been encoded using human knowledge, it has shown the possibility that simulated movements may be used for the training.

Computer vision researchers also have studied methodologies for the image composition itself. [7] developed a methodology to synthesize images and videos. In addition, a robust image composition algorithm which is able to paste video objects to background images was presented in [1]. [6] showed the potential that simple motion (e.g. running) of a person can be composed. Their system was able to synthesize animated videos from the learned motion model.

Active learning. Active learning, the learning process involving user interventions, has been popularly used for applications with limited amount of human resources [19, 15]. Recently, several computer vision researchers have adopted the active learning algorithms for recognizing objects from images. The active learning methodologies with support vector machines (SVMs) applied to the object recognition (e.g. [3]) have shown particularly successful results. In addition, [22] have suggested that active learning is able to benefit processing videos, such as recognizing people appearing in videos.

3. Video Composer

In this section, we describe our methodology to compose new activity videos based on an original video. Our video composer generates multiple videos with varying activity structures, actor locations, scales, and clothings, so that they can be used for the recognition of the activity. We first describe how we represent an activity video in terms of its components in Subsection 3.1. Re-generating new videos while modifying the components of an original video representation provides us various composed videos, which we discuss in Subsection 3.2.

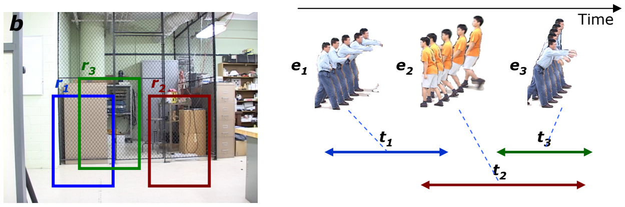

Figure 3. An example video representation of a person 'pushing' the other. The 'pushing' is composed of three sub-events, a person 'stretching' his/her arm (s_1), the person 'withdrawing' it (s_3), and the other person being pushed away (s_2). A background image, sub-event videos, spatial bounding boxes, and time intervals specifying the temporal ordering among the sub-events are presented.

3.1. Video Representation

We represent an activity video as a composition of its scene background and the foreground videos of the sub-events composing the activity. Sub-events are atomic-level actions (e.g. stretching an arm, withdrawing an arm, and moving forward), each performed by its corresponding actor in the scene. For example, a human-human interaction of 'pushing' is composed of multiple sub-events including one person 'stretching an arm' and the other person 'pushed away' as a consequence. The motivation is to maintain the foreground video segments corresponding to these sub-events, and generate videos by pasting them into the background. Essentially, we are decomposing the entire video into multiple components describing salient (and semantically important) movements, so that the video composer is able to reconstruct it by pasting them to the background.

For each sub-event, a sequence of background subtracted images (i.e. a foreground video) illustrating the actor movements is maintained. In addition, a bounding box describing 'where to paste the sub-event video' spatially and a time interval (i.e. a pair of starting time and ending time) describing 'which frames to paste the video' are associated per sub-event as its spatio-temporal region. Modifying these will change the structure of the activity, enabling us to generate various composed videos with different temporal organizations. Figure 3 shows an example video representation.

Formally, we represent an activity video V by its three components $V = (b, G, S)$. b is the background image. G describes the spatial location of the activity's center, the spatial scale, and the temporal length of the video: $G = (c, d, o)$. S is a set of sub-events, $S = \{s_1, s_2, ..., s_{|S|}\}$, where s_i is the ith sub-event. Each s_i contains four types of information: $s_i = (e_i, a_i, r_i, t_i)$. e_i is the sequence of foreground images obtained during the sub-event, $e_i =$

$e_i^0 e_i^1 ... e_i^{n_i}$ where n_i is the length of the foreground video. a_i indicates the actor id performing the sub-event. r_i is the normalized bounding box specifying the sub-event's spatial location, which is described relatively with respect to c: $r_i = (r_i^{left}, r_i^{right}, r_i^{height}, r_i^{width})$. r_i multiplied by d specifies the relative occurring region of the sub-event s_i in the video scene. $t_i = (t_i^{dur}, t_i^{loc})$ is a normalized time interval specifying the center of the interval and its duration. Let the starting time of the ith sub-event be $start_i$ and the ending time of it be end_i. Then,

$$\begin{aligned} t_i^{loc} &= \frac{start_i + end_i}{2 \cdot o} \\ t_i^{dur} &= \frac{end_i - start_i}{o}. \end{aligned} \tag{1}$$

That is, we are normalizing the occurring time and the duration of a sub-event with respect to o. Thus, the real duration of the sub-event in the video is computed as $t_i^{dur} \cdot o$.

The procedure to construct the video representation V_{ori} from an original video is straight forward. We assume that the background image b is provided or automatically estimated. In the original video, the spatial scale d is always defined to be 1. o is computed to be the number of frames of the given video. The time intervals of the sub-events are also assumed to be provided or automatically estimated, and the other parts of the representation will be obtained based on them. The system subtracts foreground regions using the given background image b, and tracks each individual appearing in the scene. The bounding box is calculated per sub-event so that the acting person is included in it spatially during the entire time period of the sub-event. c is computed by averaging the coordinates of all bounding boxes, and we subtract c from each of the boxes to get r_i. A foreground image inside the region r_i at each frame of the given time interval is concatenated to form e_i.

Once the representation is constructed, we are able to re-generate the video corresponding to the representation. Furthermore, foreground videos of the sub-events, e_i, are maintained independently in our video representation, suggesting that we are able to paste them to any desired spatio-temporal regions (r_i, t_i). This enables the composition of a video with a different temporal organization of the sub-events, generating a new video with a different activity structure.

3.2. Video Composition

In this subsection, we present a methodology to compose a video sequence from its representation. The motivation is to generate composed videos from multiple representations, Vs, produced by giving variations to the original video representation V_{ori}. We discuss how a video is composed from an arbitrary representation V with a particular background, actor location, scale, and temporal structure (i.e. b, c, d, o, r_i, and/or t_i).

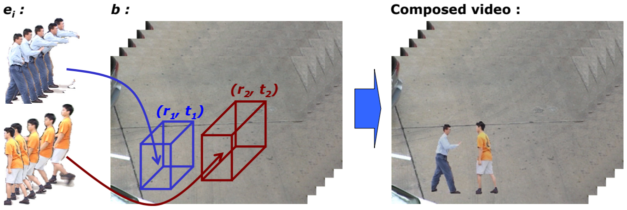

Figure 4. A video is generated from the representation by pasting each e_i to the background. The spatio-temporal location to paste the sub-event video is specified with (r_i, t_i). The pasting process of the third sub-event, a person 'withdrawing' his/her arm, is omitted here.

For any given representation V, a new video sequence is composed by pasting all sub-event videos (i.e. e_i) to the background image b. The location to paste each foreground video within the background is specified in r_i, and the frames to paste must be chosen considering t_i. In addition, the pasting process must be spatially translated by c and scaled by d, generating a video with the length o. Figure 4 illustrates an example video composition process.

More specifically, the spatial locations and the frames to paste a sub-event to generate the video for V is calculated as follows: The spatial bounding box to paste the sub-event s_i is computed by multiplying r_i by d (i.e. scaling) and then adding c to the location of the box r_i (i.e. translating).

$$box_i = d \cdot r_i + c. \tag{2}$$

The sub-event video e_i must be pasted to the frames between the frame number $start_i$ and end_i:

$$\begin{aligned} start_i &= \lfloor t_i^{loc} \cdot o - \frac{t_i^{dur} \cdot o}{2} \rfloor \\ end_i &= \lfloor t_i^{loc} \cdot o + \frac{t_i^{dur} \cdot o}{2} \rfloor. \end{aligned} \tag{3}$$

For each sub-event s_i, the video composer pastes the frame e_i^j of the sub-event video to the kth frame of the video being composed. That is, for every frame k where $start_i \leq k \leq end_i$, we calculate the corresponding frame j of the sub-event video while considering the overall duration of the sub-event t_i^{dur}.

$$j = \lfloor \frac{(k - start_i) \cdot n_i}{t_i^{dur} \cdot o} \rfloor. \tag{4}$$

In addition, the video composer pastes the actor for frames between sub-events. The assumption is that the actor is staying stationary if no sub-event is performed, since all his/her salient movements have already been encoded as sub-events. For each actor, the video composer estimates his/her appearance in every frame l that has not been covered by any sub-event. Basically, it searches for the temporally nearest sub-event s_q, and assumes that the appearance of the actor is identical to that of the closest frame of the sub-event. That is, we paste $e_q^{n_q}$ to frame l if end_q is less than l, and paste e_q^0 otherwise.

Furthermore, our system also supports various image operations such as flipping and color changing, in order to increase the diversity among composed videos. The video composer is able to flip the foregrounds being pasted, if necessary. Also, color changes of actors' clothings are supported: A color clustering algorithm is applied to e_i to detect and track color blobs of the actors. Based on the tracked blobs, our video composer changes the color intensities of the upper-body blobs and the lower-body blobs randomly. An overall intensity of the foreground is also adjusted considering the background illumination.

4. Active Structure Learning

In this section, we present a methodology to learn a decision boundary distinguishing correct activity representations from abnormal representations. In principle, the video composer presented in the previous section is able to generate a video from any given representation V. However, not every temporal structure, which essentially is a set of sub-events' time intervals, is possible for the activity. For example, in the case of 'shaking hands', the timing of the sub-event 'withdrawing an arm' of one person must not occur before the other person 'stretching an arm'. If this is violated, the representation must be treated as another activity or irrelevant noise. The system must estimate the decision boundary specifying which temporal structure is valid for the activity, so that it is able to compose videos only from the representations within the boundary.

One of the major difficulties in estimating the decision boundaries is that the system has a limited amount of training examples: The system only has a single positive example, V_{ori}, and few trivial negative examples are known (e.g. a representation with all time intervals' durations set to 0). In order to overcome such difficulty, we have designed an active learning algorithm that takes advantage of the video composer described in the previous section. Our algorithm 'actively' analyzes the structure space by generating necessary proposal videos having various structures. The decision boundary is updated iteratively based on the labels of the proposal videos, which is obtained from an oracle (i.e. a human). Videos are generated and labeled based on their necessity, in contrast to previous passive learning methodologies.

We model a video representation's temporal structure as a vector concatenating all sub-events' time intervals as well as the activity's overall speed. That is, we concatenate o and all t_i to form a vector x with length $2 \cdot |S| + 1$. In this $2 \cdot |S| + 1$ dimensional space, the goal is to construct a classifier deciding whether a given vector x corresponds to the correct structure of the activity or not.

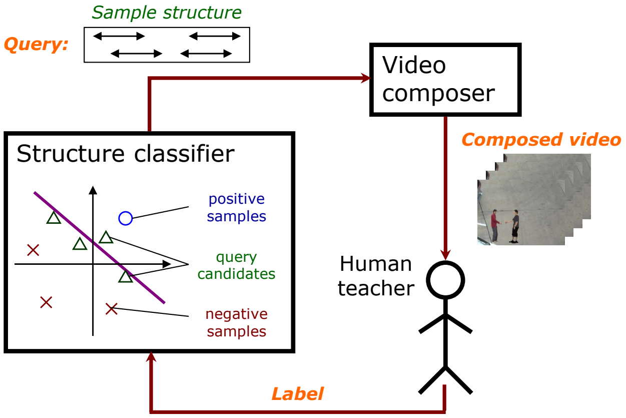

Figure 5. An overall learning process of our system. At each iteration, a query (i.e. a temporal structure) is given to the video composer to generate a composed video. The video is given to an oracle (i.e. a human) to obtain the label specifying whether the generated video contains the valid activity or not. The classifier is updated based on the label, for the next iteration.

The overall flow of our active learning algorithm is as follows. At each iteration, the system calculates the decision boundary based on the labeled structures (i.e. vectors) it already has. Next, the system randomly samples several proposal video structures, x_m, and choose one among them which is believed to be the most informative. A video is composed based on the chosen vector x_{min}, by modifying V_{ori}. Then, the system requests the label of the composed video (whether the video contains a correct activity or not) from a human teacher. That is, it generates a single query per iteration. The labeled vector is provided to the system as a new example, leading to the next iteration which will calculate a new decision boundary. Figure 5 illustrates such process.

We use a support vector machine (SVM) classifier for our active learning. A SVM classifier divides the entire space with a hyperplane (i.e. the decision boundary). In our SVM-based active learning, the data point nearest to the decision boundary is assumed to be the most informative, similar to [15]. This heuristic have been confirmed to be computationally efficient (only a dot product computation is required) and effective.

Let $w \cdot x + a = 0$ be the hyperplane separating valid and non-valid structures. Our system searches for the vector x_{min} minimizing the distance between the vector and the hyperplane:

$$x_{min} = argmin_{x_m} \frac{w \cdot x_m + a}{||w||} \tag{5}$$

where each x_m is a uniformly sampled vector. The algo-

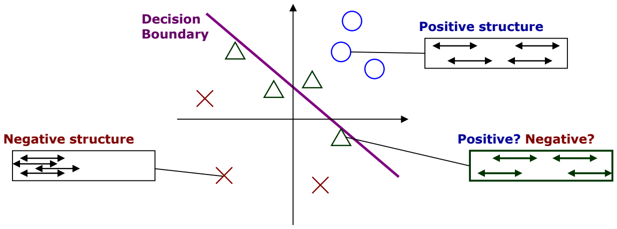

Figure 6. An example decision boundary and video structure samples. Blue circles indicate structures with positive labels (i.e. they are valid structures), and red crosses indicate invalid structures. Purple line is a hyperplane separating valid and invalid structures (i.e. the decision boundary). Green triangles are the candidate structures that have not been labeled. The goal is to choose and label one of those ambiguous structures (i.e. green triangles), to update the decision boundary for the next iteration.

rithm chooses x_{min}, the most informative vector among randomly sampled vectors. The SVM classifier is updated by labeling x_{min}. Figure 6 shows example vectors and a SVM boundary.

Any video representation with the structure sampled inside the learned decision boundary will produce a valid activity video satisfying the constraints. This enables random composition of multiple activity videos with varying structures.

5. Adaptive Training Set Construction

This section describes a methodology to automatically construct a set of composed training videos for the activity recognition. Using the methods from the previous sections, we compose videos with various sub-event locations, scales, and temporal structures while making their backgrounds to be identical to the testing environment. The idea is to compose activity videos that are expected to be observed from the current scene, so that the training videos tailored for the scene are generated.

We generate training videos by randomly sampling possible representations. The detailed procedure for our training set construction is as follows. Given the original video representation V_{ori}, we make a new video representation V while copying the sub-event videos e_i and their spatial locations r_i from V_{ori}. As discussed above, the background image b of V is set to the current background (i.e. testing) image. The temporal structure (i.e. o and t_i) of the representation is randomly selected while discarding the abnormal structure: The learned SVM classifier (i.e. Section 4) is able to judge whether the temporal structure will form a correct activity video or not. Only the video representations that satisfy the learned activity structure will be sampled, and their videos will be generated. The location c and scale d of the video V are also chosen randomly. Image operations (e.g. flipping) mentioned in Subsection 3.2 are randomly applied as well, increasing the diversity. This process of creating new representation V and composing a video from it is repeated multiple times, to construct a large training set.

As a result, a set of training videos with various appearance and motion is produced for each activity. In principle, our approach presented throughout the paper is able to support any activity recognition methodology by providing the constructed training set. That is, our approach for solving the problem of one-example activity recognition is to provide a new training set with abundant composed examples, making our active video composition applicable to various recognition systems.

6. Experiments

Here, we evaluate our recognition approach utilizing the active video composition. We confirm the benefits of our approach by implementing several activity recognition methodologies with and without our video composition. When training the systems, only a single training video taken from a different background with different actors is provided per activity. The details of the system implementation is presented in Subsection 6.1, and the results are discussed in Subsection 6.2.

6.1. Implementation

In order to measure the effectiveness of our active video composition approach, we constructed various activity recognition systems with and without our video composer. Multiple classifiers using different types of features have been implemented, and they have been trained with a set of composed videos or only with a set of original videos.

Each implemented system extracts one of the two types of spatio-temporal local features, [5] and [8], to recognize activities. These features have been confirmed to be robust to background changes and camera movements. We apply a feature extractor to each 3-D volume constructed by concatenating video's image frames along time axis, obtaining a set of 3-D XYT points with salient appearance changes. Next, a feature point is categorized into multiple types (e.g. 400) based on the appearance of the local 3-D volume around it. As a result, each system constructs a histogram of feature appearance types per video, which essentially is a feature codebook.

The systems categorize their codebooks into classes of human activities. Various classifiers have been implemented and tested, including basic K-nearest neighbor classifiers and SVM classifiers. In the basic K-nearest neighbor classifier, the similarity between a testing video and a composed training video is measured by computing the linear distance between between two codebooks. The recognition is performed by examining the labels of the K

most similar training videos. SVM classifiers which estimates hyperplanes separating activity classes have also been implemented. In addition, we have implemented a non-hierarchical version of the spatio-temporal relationship match (STR match) [13]: The implemented STR match measures structural similarity between two videos by analyzing feature relations while using SVM classifiers.

As a result, a total of 6 systems (listed in Table 1) have been implemented and tested with/without our video composition. The reason why we have chosen discriminative classifiers is that they inherently possess an ability to perform the classification from few examples (e.g. not many training examples are required to construct a K-NN or SVM classifier), enabling us to test the methodologies even with a single video. These systems implemented in our experiment represent the state-of-the-art activity recognition approaches using spatio-temporal features.

6.2. Evaluation

We have tested the performances of the systems on a high-level human activity recognition task. Videos of six types of human-human interactions, shaking hands, hugging, kicking, pointing, punching, and pushing, are classified using the systems. These interactions are complex human activities performed by multiple actors, and their recognition is a challenging problem especially when a limited amount of training videos are provided. Furthermore, most of these interactions share common movements such as 'stretching' and 'withdrawing', preventing the classification. For our experiments, we have used two different existing datasets with distinct characteristics, one for the testing and one for the training. The objective is to emulate real world scenarios where training and testing videos have different properties (e.g. backgrounds).

We used the UT-Interaction dataset #1 from the SDHA 2010 activity recognition contest [14] as our test set, which was also used in [13]. This public dataset consists of videos of the above-mentioned six types of interactions, containing 10 executions per activity. Videos were taken with 10 different settings, each having different background, scale, illumination, and actors. In addition, irrelevant pedestrians are present in the videos. In our experiment, we have used the segmented version of the dataset where each video segment contains only one activity execution. Video regions are cropped spatially and temporally, giving us a total of 60 various-sized videos with 15 fps. The entire videos have been used for the testing.

As an original training video, we have selected one sequence per activity from the dataset of [12]. A label indicating the activity type, its spatial location, and its time interval are provided per video. Further, sub-events composing the activities have been labeled, so that the system is able to compute the representation of the video. Each activity is decomposed into 2 to 4 different atomic-level actions organized sequentially and concurrently, such as a person 'stretching' an arm and a person 'moving' forward. Background images of the training videos have been provided as well. The representations have been constructed automatically based on the backgrounds and the sub-event labels, following the algorithm mentioned in 3.1.

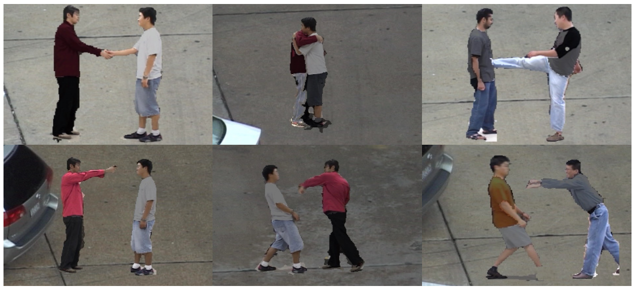

Figure 7. Example snapshots of composed training videos. The figure only shows the activity regions cropped from the entire videos.

For the video composition, the active learning algorithm of Section 4 has been applied to train the SVM classifier modeling each activity's structure. Each SVM classifier has been trained with 20 iterations (i.e. 20 queries were given to a human sequentially, obtaining their labels). The composed training set has been created by generating activity videos with rough background image segments of the test set (Figure 7).

Table 1 compares the activity *classification* accuracies of the systems. Each system has been tested with four different training settings: using a single real training video per activity, using 10 real training videos from [12] per activity, using 10 composed videos generated from a single real video per activity, and using 30 composed videos also generated from a single real video per activity. The result confirms the benefits of our active video composition: The systems with composed videos performed much better than the systems without them (e.g. 0.317 vs. 0.533). Even though the backgrounds, locations, scales, actor clothings, and their movements in the testing videos were very different from those in the training video, our approach with the video composer was able to recognize human activities reliably.

In addition, we are able to observe that the performances of the systems with 10 composed videos are higher than those with 10 real videos, even though only one real video was used to generate the composed videos. This is due to our method's ability to compose training videos tailored for the testing environment (e.g. background). Our video composer takes advantage of its knowledge on testing environments such as lighting conditions and backgrounds, generating training videos much more similar to the testing videos then the original videos.

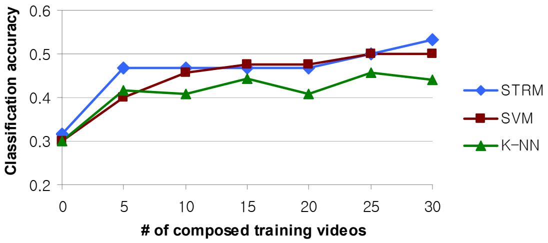

Figure 8. Accuracies of the systems with respect to the number of composed videos. We only present results of the systems using [8]'s features. Those with [5] showed similar behavior. The performances have been averaged for four training set constructions.

Table 1. Activity classification accuracies of the systems tested on the UT-Interaction dataset #1 [14].

Systems	1 real	10 real	10 comp.	**30 comp.**
Random chance	0.167	0.167	0.167	0.167
[5] + K-NN	0.300	0.317	0.417	0.433
[5] + SVM	0.300	0.383	0.433	0.450
[5] + STR match	0.333	0.433	0.450	0.467
[8] + K-NN	0.300	0.350	0.400	0.450
[8] + SVM	0.300	0.383	0.450	0.500
[8] + STR match	0.317	0.433	0.467	**0.533**

We also have conducted the experiments to measure the recognition accuracies with respect to the number of composed training videos used per activity. Figure 8 shows the experimental results. The increase in the number of composed videos benefits the system performances, justifying our overall paradigm of generating various composed videos. The video composition itself runs in real-time with our unoptimized C++ codes (Intel i7 940 CPU used). The recognition was also performed in real-time, except for the adopted feature extraction part.

7. Conclusions

We have presented a new approach to recognize a human activity using a single video example. We introduced the paradigm of utilizing composed videos for the recognition, which are generated by manipulating an original video. A methodology to construct representations describing activity videos, and that to generate composed videos from the representations have been discussed. In addition, an active learning algorithm to learn the structure of the activity has been designed. As a result, our system generated a training set of composed activity videos tailored for the given environment. Experimental results confirmed that our approach benefits the recognition of complex human activities by providing diverse composed videos. In the future, we plan to extend our active video composition to take advantage of training videos created with fully synthetic 3-D agents. Creating activity videos with CG animations will enable us to handle 3-D actor/camera movements, recognizing a wider range of human activities.

Ack.: This work was supported partly by the R&D program of the Korea Ministry of Knowledge and Economy (MKE) and the Korea Evaluation Institute of Industrial Technology (KEIT). [The Development of Low-cost Autonomous Navigation Systems for a Robot Vehicle in Urban Environment, 10035354]

References

[1] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: robust video object cutout using localized classifiers. In *ACM SIGGRAPH*, 2009.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005.

[3] E. Y. Chang, S. Tong, K. Goh, , and C. Chang. Support vector machine concept-dependent active learning for image retrieval. *IEEE Transactions on Multimedia*, 2005.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research (JAIR)*, 16:321–357, 2002.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Workshop on VS-PETS*, pages 65–72, Oct 2005.

[6] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, volume 2, pages 726–733, Oct 2003.

[7] V. Kwatra, A. Scholdl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: Image and video synthesis using graph cuts. In *ACM SIGGRAPH*, 2003.

[8] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[10] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3), Sep 2008.

[11] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE T PAMI*, 22(8):831–843, 2000.

[12] M. S. Ryoo and J. K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. *IJCV*, 32(1):1–24, 2009.

[13] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

[14] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846, 2000.

[15] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.

[16] H. J. Seo and P. Milanfar. Detection of human actions from a single example. In *ICCV*, 2009.

[17] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007.

[18] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *ICML*, 2000.

[19] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE T CSVT*, 18(11):1473–1488, Nov 2008.

[20] S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *CVPR*, 2007.

[21] R. Yan, J. Yang, and A. Hauptmann. Automatically labeling video data using multi-class active learning. In *ICCV*, 2003.