Multiple Scale-specific Representations for Improved Human Action Recognition

Amir H. Shabani[a,b], John S. Zelek[b], David A. Clausi[a]

[a] *Vision and Image Processing (VIP) Lab*
[b] *Intelligent Systems Lab*
Department of Systems Design Engineering
University of Waterloo, Waterloo, ON, Canada, N2L 3G1
E-mails: hshabani,jzelek,dclausi@uwaterloo.ca

Abstract

Human action recognition in video is important in many computer vision applications such as automated surveillance. Human actions can be compactly encoded using a sparse set of local spatio-temporal salient features at different scales. The existing bottom-up methods construct a single dictionary of action primitives from the joint features of all scales and hence, a single action representation. This representation cannot fully exploit the complementary characteristics of the motions across different scales. To address this problem, we introduce the concept of learning multiple dictionaries of action primitives at different resolutions and consequently, multiple scale-specific representations for a given video sample. Using a decoupled fusion of multiple representations, we improved the human classification accuracy of realistic benchmark databases by about 5%, compared with the state-of-the art methods.

Keywords: Human action recognition, scale-specific representation, concatenated representation, decoupled representation, spatio-temporal salient features, separability test.

1. Introduction

Humans can easily detect and recognize the type of actions performed in a video. However, the automatic recognition of human actions [1, 2, 3, 4] is a challenge in computer vision with growing applications for automated surveillance [5], content-based video retrieval [6], video summarization [7], elderly home monitoring for assisted living [8], and human-computer interaction [5]. The confusion lies in people performing the same action in noticeably different ways, leading to errors of omission. Also, individuals performing different actions that visually appear to be similar, lead to errors of commission. In addition, illumination and view/scale changes create further challenges to automatically interpret the scene.

The discriminative bottom-up approaches become more popular for human action recognition in an unconstrained setting such as youtube videos. A widely used approach is a bag-of-words (BOW) [1, 2, 4, 9] framework (Fig. 1(a)) in which the video contents are sparsely localized by the salient changes such as starts/stops of subactions. In this framework, the salient features are first extracted at multiple spatial and temporal scales. A single dictionary of action primitives (i.e., visual words) is then learnt from the joint features of all scales from the training video samples. Conventionally, an action is represented by a normalized histogram which shows the frequency of the multi-scale features over the action primitives. Finally, a support vector machine (SVM) classifier with a matching kernel such as linear, χ^2, or (Gaussian) radial basis function [4, 9, 10, 11] categorizes an unknown action representation according to its distance from the learnt decision boundaries during training.

There are three main elements in a BOW framework which directly affect the final action classification accuracy: (1) the quality of salient features which capture the local video events, (2) the descriptiveness of the dictionary of action primitives, and consequently, the discriminant of the actions representations, and (3) the matching strategy and type of classifier. Different methods use different quality features with different classifiers [1, 2, 4, 9, 11], but most of these methods use a single dictionary of action primitives and a single action representation which cannot fully exploit the complementary characteristics of the motions at different scales and hence, this single dictionary is not sufficiently robust to represent accurately all different motion patterns. Moreover, the intrinsic scale from which the salient features are extracted is a discriminant information which cannot be encoded in the single action representation.

This paper proposes two alternatives to the single non scale-specific dictionary learning and hence, the single action representation to improve the discrimination of different actions and consequently, to boost the classification accuracy. To address the limitations of a single action representation, we propose to learn a separate dictionary of action primitives for each individual scale and analyze the features of each scale independently. A distinct representation of an action is then obtained using the salient features extracted at a given spatio-temporal scale encoded by the corresponding dictionary. We will thus have multiple representations of the same action at different scales in which the intrinsic scale of the features are accordingly encoded by construction. There are two viable approaches to fuse these

multiple representations: concatenated and decoupled. In the former approach, the representations are concatenated in scale order into one long vector. In a decoupled approach, the representations are kept separate until their matching at the classification stage according to their relative discriminant importance. Note that the concatenation approach is a typical policy for combining multiple vectors [12] and the decoupled approach is our introduced fusion approach. Both approaches are viable methods and we thus consider both for performance comparison in this paper.

The main contribution of this paper is the introduction of multiple scale-specific dictionaries with a concatenated and decoupled action representations which benefit from the complement and discriminant motion information among multi-scale salient features. This choice is motivated by the benefit of complementary motions in multi-scale features and multiple dictionaries of action primitives, the redundancy in the dense scale sampling [3], the elimination of heuristics in intrinsic features' scale selection with inherent artifacts [11], and the usefulness of multiresolution histograms [13]. We also introduce a quantitative measure to evaluate the separability of different action representations and examine their performance comparison in a common classification framework on different benchmark human action recognition datasets.

The rest of this paper is organized as follows. Section 2 reviews the use of local salient features for action representation in the existing methods. Section 3 explains the concept of learning multiple dictionaries of action primitives and representations. Section 4 describes the experimental setting. Section 5 introduces the separability test to compare different representations. Section 6 presents the action classification accuracy of using different representations on different datasets. Finally, Section 7 concludes the paper.

2. Literature Review

Most of existing action recognition methods either use a top-down approach or a bottom-up approach. A typical top-down approach requires foreground segmentation by using a shape or appearance model of a human for detection and tracking [14]. In a constrained environment, the top-down methods have shown promising results [5], but the segmentation and tracking might not be reliable in the presence of background/camera motion, clutter, or occlusion in a real-world low resolution videos. A typical bottom-up approach [1, 2, 4], on the other hand, can learn a set of action primitives using local spatio-temporal salient features without needing to perform video segmentation and tracking. This paper focuses on this second methods with a discriminative approach.

A local salient feature [1, 3, 4, 15] is centered at a spatio-temporal key point whose saliency score such as cornerness [15] or motion energy [1, 4] is the highest in its local spatio-temporal neighbors. The feature is described by the shape and/or motion characteristic of the pixels in its distinct volume and has shown to be more robust to clutter and occlusion [5] when compared with the global features. Extracted at multiple spatial and temporal scales, the local salient features can provide a sparse and compact representation of the video contents. In the next subsections, we first explain the existing salient feature extraction methods and the standard representation of actions using these features. We then highlight the limitations of the standard single action representations and introduce our contributions addressing these limitations.

2.1. Spatio-temporal salient feature extraction

The existing local salient spatio-temporal features can be categorized into two groups: structured-based and motion-based [9, 16]. By treating a 2D+t image sequence as a 3D object, structured-based feature detectors use the same type of filters in the spatial and temporal directions to detect 3D salient structural features such as 3D Harris corners [2] or 3D Hessian ellipsoids [3]. Motion-based salient feature detectors such as Cuboids [1] consider different spatial and temporal filtering such as 2D spatial Gaussian and a temporal Gabor filtering. The motion-based features have shown to be more effective than the structured-base features for action recognition [16].

Recently, asymmetric temporal filters such as Poisson, truncated exponential, and asymmetric sinc have shown to be more effective than the widely used symmetric temporal filters such as Gaussian or Gabor for salient motion feature detection [4, 10, 17]. The features detected using a temporal asymmetric, complex sinc filter has shown the highest robustness under different geometric deformations and provided the highest classification accuracy [4]. We refer to these features as asymmetric motion features throughout this paper and use them in our experiments.

2.2. Standard single action representation

In the standard BOW framework (Fig. 1(a)), the salient features detected at multiple spatio-temporal scales from all training samples are combined to learn a single action dictionary/codebook. The elements of this dictionary are referred to as primitives, attributes, prototypes, or visual words [1, 2, 3, 4, 9] and we will use the term action primitives. Conventionally, for the dictionary learning using a standard vector quantization (VQ) [18], the k-means algorithm groups the features with similar motion and appearance patterns in the same cluster referred to as a visual word. The features are assigned to the closest cluster for encoding their corresponding video contents. An action is then described by a single representation $\mathbf{x}_S$ which is the L_1-normalized term frequency occurrence of the features in the whole video over the dictionary of action primitives [1].

The standard single action representation provides a global (without context) representation of the video contents. To add some contextual information, the multi-channel SVM [11] uses the relative spatial and temporal localization of the salient features by representing the video contents at multiple channels (i.e., different divisions of the video volume). In this approach, different SVMs are trained using different combination of the channels and the best channel combination is chosen by cross

validation or using a validation set. The multi-channel SVM approach still uses a single dictionary of action primitives. Moreover, the spatio-temporal extension of each channel, the number of channels, and the best channel selection approach are heuristic, data dependent, and computationally expensive at the training stage.

2.3. Limitations of a single action representation

There are two problems with learning a single dictionary of action primitives from the joint features of multiple scales and consequently, a single action representation. (1) The discrimination power of the multi-scale features is lost due to early stage fusion of the features before dictionary learning. In fact, the salient features at different spatio-temporal scales encode different motion characteristics and might have complementary and/or redundant motion information. More specifically, the coarse-scale features are more descriptive of average motions while the fine-scale features capture the details of the motions. For example, actions such as running and jogging have a similar average motion pattern with potentially similar coarse-scale features. The fine-scale features should thus provide a better discrimination of these motions (we provide more justifications on this argument in Section 5). (2) The salient features are extracted at different spatio-temporal scales. Encoding the intrinsic scale of the features provides a key discriminatory information to differentiate motion patterns of different actions. This information cannot be incorporated in the single non-scale-specific action representation. This is due to the fact that the dictionary of action primitives is general for the features of all scales (we provide two approaches in Section 3 to encode this information).

2.4. Contributions

This paper has two main contributions to improve the representation of different actions and consequently, their classification.

1. To construct a more discriminant action representation and to fully benefit from the complementary motion information in the set of multi-scale features, we propose to learn multiple dictionaries of action primitives across different spatio-temporal scales and compute the corresponding scale-specific representations of an action (Section 3.1). As different dictionaries might provide a complementary set of action primitives, using multiple action representations should improve the discrimination power of the action representations and hence, the classification accuracy.
2. We propose two improved, novel action representations in which the intrinsic scale of the features can be encoded in the representation. We also introduce a specific method to incorporate the discriminant level of different representations in the matching of a test video with the training samples at the classification stage (Sections 3.2 and 3.3).

2.5. Research questions

We highlight three research questions regarding the evaluation of the above-mentioned contributions and investigate the discrimination of scale-specific representations and action classification in video.

1. How does the discrimination of action representations across different scales changes? This question is addressed using separability test in Section 5.1.
2. How does encoding multi-scale features using scale-specific dictionaries affect discrimination of action representation. Separability tests are performed in Section 5.2 to answer this question.
3. How successful are the multi scale-specific representations for action classification? And also, between concatenated and decoupled fusion approach, which one performs better. This research question is examined in Section 6.

3. Multiple Scale-specific Action Representations

In this section, we explain our contributions in constructing multiple scale-specific action representations from different dictionaries of action primitives. We also introduce two different strategies in fusion and matching of these representations.

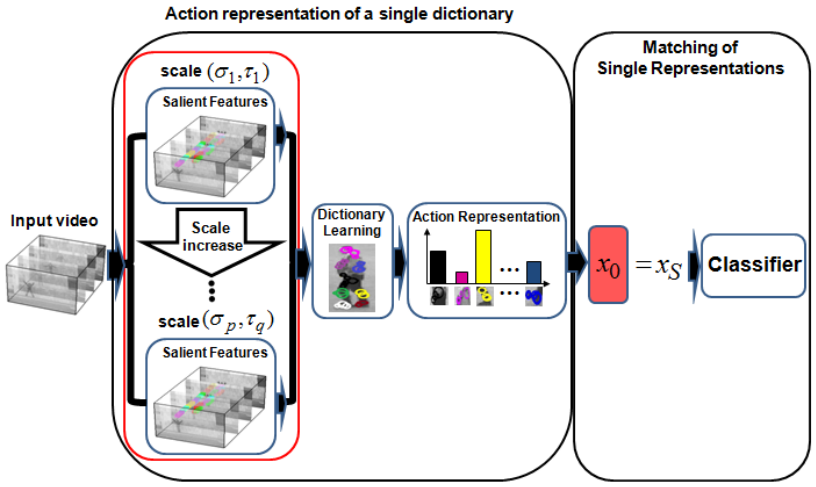
3.1. Scale-specific dictionaries and representations

To fully exploit the multiresolution characteristics of the motion patterns, we propose to learn a dictionary of action primitives for each individual spatio-temporal scale and consequently, describe an action at different scales (Fig. 1(b) and 1(c)). In this treatment, the features at each specific scale are processed independently and each scale-specific action representation is constructed from the corresponding dictionary and features of that specific scale. Our hypothesis is that such multiple representations should incorporate more discriminatory information in the action representation which consequently should boost the classification accuracy. Having multiple representations of an action, an effective fusion strategy helps the classifier to fully benefit from these complementary representations.

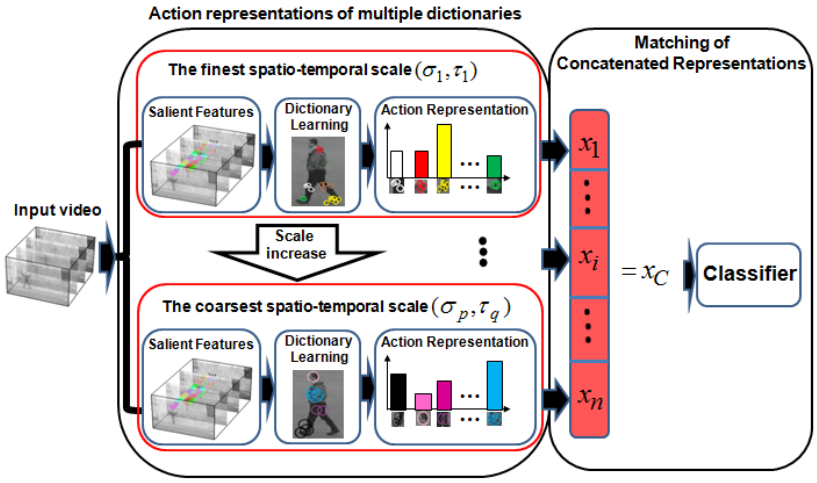
In the next subsections, we introduce two different strategies to fuse the multiple scale-specific action representations with their corresponding scales: concatenation (Section 3.2) and decoupled (Section 3.3). Before proceeding further, note that the standard representation is a single non-scale-specific representation. Our alternatives are a concatenated or decoupled multi scale-specific representation. For conciseness, we will refer to these representations as single, concatenated, and decoupled.

3.2. Concatenated action representation

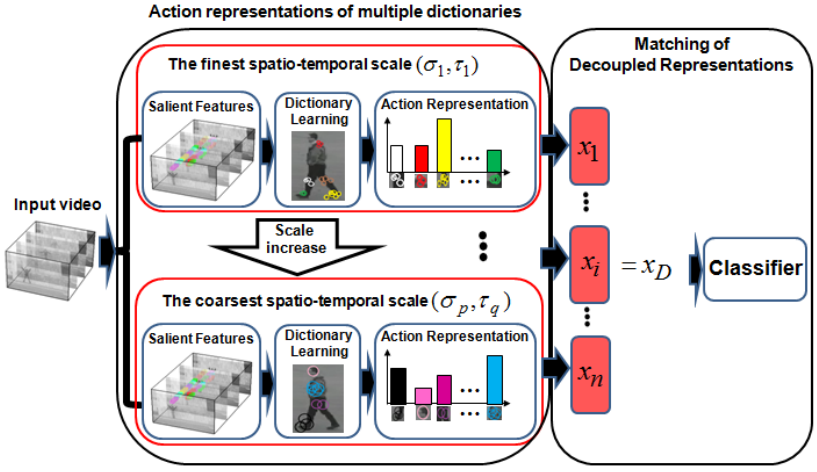
Assume $\mathbf{x}_1$ denotes a representation of an action using the finest-scale features and their corresponding dictionary of action primitives. One way of fusing representations of an action at different spatio-temporal scales $i = \{1, 2, ..., n\}$ is to concatenated them in order to form the final representation vector



(a) Standard approach: a single action representation



(b) New approach: a concatenated action representation



(c) New approach: a decoupled action representation

Figure 1: (a) In standard bottom-up discriminative action recognition [1, 2, 4, 9], the multi-scale features are combined to learn a single dictionary of action primitives and a single action representation ($\mathbf{x}_S$). To fully exploit the multiresolution characteristics of these features at different spatio-temporal scales, we propose to learn a separate dictionary and hence, a separate action representation $\mathbf{x}_i$ at each scale $i = \{1, 2, .., n\}$. We can then describe an action as (b) concatenated representation ($\mathbf{x}_C = [\mathbf{x}_1, ..., \mathbf{x}_n]$) in which all the representations are concatenated in one large vector or as (c) decoupled representation ($\mathbf{x}_D = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$) which keeps all the representations in a set to incorporate the order of importance of different representations in the matching of a test video with the training samples.

$\mathbf{x}_C = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ to be fed to a classifier (Fig. 1(b)). For the case in which the action representations of each individual scale is a d-dimensional vector and we have n spatio-temporal scales, the concatenated multiple representation is a $(n*d)$-dimensional vector. In a general case, the length of the dictionary at each scale might be different and hence, the action representations might have different dimensions.

The concatenation of the action representations in one feature vector to be fed into a classifier might be problematic due to the curse of dimensionality [19, 20, 21, 12] with increase in the dimensionality of each representation or the number of spatio-temporal scales. In fact, when the dimensionality increases, the volume of the space increases so fast that all representations appear to be sparse which prevents the classifier from being efficient. This problem is more severe when the number of training samples is small relative to the dimensionality of the action representation [22].

In the next subsection, we propose matching of the decoupled action representation to overcome on the curse of dimensionality problem and also incorporate the prior information on discriminant level of the representations at different scales.

3.3. Decoupled action representation

Concatenation has two main problems. (1) The curse of dimensionality can reduce the efficiency of the classifier. (2) Action representations at different scales have different discrimination power. The concatenation fusion approach cannot utilize this useful prior information at the matching stage. Intuitively, a finer-scale representation should better capture the statistics of the video contents than a coarser-scale representation. This is due to the fact that the salient features detected at the finer scales are less dislocated (due to spatial Gaussian smoothing) [23, 24], more precise, and more robust than those detected at the coarser scales [4]. The corresponding finer-scale dictionary of action primitives and representation should thus be more descriptive and more discriminant (we perform separability tests in Section 5 to validate this idea). This domain knowledge prior information should be explicitly incorporated in matching to improve the classification accuracy. To address the above mentioned problems of a concatenated representation, the multiple representations should be kept decoupled until the matching at the classification stage (Fig. 1(c)). To differentiate from the concatenated approach in which the representations are stacked in order and form a single vector $\mathbf{x}_C$, we represent the decoupled multiple representations as a set $\mathbf{x}_D = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$.

Constructed as a set, a decoupled representation requires a specific way to be compared with another representation. We consider the distance between two decoupled action representations to be the sum of the distances between the best joint matched representations for each representation in the set. That is, the action representation $\mathbf{x}_i$ at a given scale $i = \{1, 2, ..., n\}$ is compared with all scale-specific action representations of a training sample and the one with the least distance is considered as the best match. This procedure starts with the finest scale and the best matched representation is excluded from the list as we proceed to the coarsest scale. This procedure favors the finer-scale representations as they are more discriminant than the coarse-scale representations. Algorithm 1 describes this procedure for computing the distance between two decoupled representations. Modeling the level of descriptiveness of different scale-specific representations in the matching of a decoupled representation should improve the matching performance and hence, the classification accuracy.

Algorithm 1 Pseudo-code for computing the distance between two decoupled action representations $\mathbf{x}_D^1, \mathbf{x}_D^2$.

Inputs: $\mathbf{x}_D^1 = \{\mathbf{x}_1^1, ..., \mathbf{x}_n^1\}$ and $\mathbf{x}_D^2 = \{\mathbf{x}_1^2, ..., \mathbf{x}_m^2\}$.
Output: $\phi(\mathbf{x}_D^1, \mathbf{x}_D^2)$
Reset the distance value:
$\phi = 0$
for $i = 1 : n$ **do**
  Find the best matched representation for $\mathbf{x}_i^1$:
  $\mathbf{x}_{j^*}^2 = \arg\min_{\mathbf{x}_j^2 \in \mathbf{x}_D^2} \varphi(\mathbf{x}_i^1, \mathbf{x}_j^2)$.
  Compute the distance:
  $\varphi_i = \varphi(\mathbf{x}_i^1, \mathbf{x}_{j^*}^2)$.
  Remove the best matched representation $\mathbf{x}_{j^*}^2$ from $\mathbf{x}_D^2$:
  $\mathbf{x}_{j^*}^2 \notin \mathbf{x}_D^2$.
  $\phi = \phi + \varphi_i$
end for

4. Testing setup

This section describes the utilized datasets, the feature extraction method, and the discriminant criterion to evaluate the separability of different actions in a representation. We also explain our classification method using a SVM classifier.

4.1. Datasets

Five benchmark human action recognition datasets have been used for the performance evaluation of different action representation methods.

The **KTH data set** [25] consists of six actions (running, boxing, walking, jogging, hand waving, and hand clapping) with 600 choreographed video samples. Twenty-five different subjects perform each action in four different scenarios: indoors, outdoors, outdoors with scale change (fast zooming in/out) and outdoors with different clothes. Each clip lasts between 10 to 15 seconds and is sampled at 25Hz with an image frame size of 160×120. According to the initial citation [25], the video samples are divided into a test set (9 subjects: 2,3,5,6,7,8,9,10, and 22) and a training set (the remaining 16 subjects). We thus use the same training/testing protocol for the classification test.

The **UCF Sports dataset** [26] includes 10 different action classes such as diving, golf swing, kicking, lifting, riding horse, run, skate boarding, walk, swing on the pommel horse, and swing at the high bar with 150 video samples collected from the Youtube website. This dataset is challenging due to diverse ranges of views and scene changes with moving camera, clutter, and partial occlusion. A horizontally flipped version of

each video is also used during training to increase the data samples [9]. We use leave-one-out (without considering the flipped samples for testing) protocol.

The **UCF Youtube dataset** [27] contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. For each category, the videos are grouped into 25 groups with more than 4 action clips in each. The video clips in the same group share some common features, such as the same actor, similar background, similar viewpoint, and so on. We thus use leave-one-group-out protocol for the classification test. That is, when learning the decision boundary of the SVM classifier, action samples of the same group will be removed all at once.

The **HOHA dataset** [28] consists of eight human actions (answer phone, get out car, hand shake, hug a person, kiss, sit down, sit up, and stand up) from 32 Hollywood movies. The dataset is divided into a test set obtained from 20 movies and the (clean) training set obtained from 12 movies different from the test set. There are 219 sample videos in the training set and 211 samples in the test set. We therefore use a training/testing protocol for the classification test.

The **HOHA2 dataset** [11] consists of twelve human actions from 69 different Hollywood movies. In addition to the eight action classes from the HOHA dataset, four classes of eating, fight with a person , get out of car, and hand shake have been added to the new dataset. The total length of the action samples is about 600k frames or 7 hours of video. The dataset is divided into a training set with 823 samples and a test set with 884 samples (in total, 1707 action samples). Training and test samples are obtained from different movies. We therefore use a training/testing protocol for the classification test. Note that HOHA2 dataset is the extension of HOHA dataset with more training and testing data samples and more classes. Moreover, the video samples in these datasets have different image resolution and the video clips in HOHA2 last longer with motion patterns unrelated to the labeled action. This makes HOHA2 much more challenging than HOHA dataset. For completeness, we use both datasets in our experiments.

4.2. Spatio-temporal salient motion features

To localize the video events of different scales, we generate the salient asymmetric motion features (i.e., key points) at nine different spatio-temporal scales (σ_p, τ_q) in which σ is in pixels and τ in frames. The scales are each generated according to $2\,(\sqrt{2})^p$ and $2\,(\sqrt{2})^q$ with $p, q \in \{0, 1, 2\}$ [4, 29]. The spatial filter is a 2D Gaussian filter and the temporal filter is the complex, asymmetric sinc filtering [4] which has shown to be the best for robust asymmetric motion feature detection. Each salient feature has a volumetric extension for its description. We used the 3D SIFT descriptor which has shown good performance in encoding motion and appearance [30]. This descriptor computes a normalized histogram of the spatio-temporal oriented gradients of the pixels inside the volumetric extension of a salient feature.

Fig. 2 shows the 2D projection of the asymmetric motion features on different datasets. As can be seen, the features are mainly from the foreground and capture the relevant moving limbs on the subjects (see [4] for more results on the precision and robustness of these features).

4.3. Dictionary learning and action representation

The standard single action representation (Fig. 1(a)) is obtained over just one dictionary of action primitives which is learnt from the joint features from all spatio-temporal scales, irrespective to their scales [1, 2, 4]. The action representation is then the normalized frequency of the multi-scale features over this dictionary.

For the concatenated and decoupled action representations, we perform separate vector quantization of the features at each scale (Fig. 1(b) and 1(c)). That is, the features of each scale are experimentally quantized into $d = 1000$ clusters for which we performed the K-means clustering with random seed initialization ten times and kept the result with the lowest error [9]. The clusters (i.e., visual words) represent the action primitives. Having nine spatio-temporal scales, we have nine different dictionaries of action primitives for describing an action using the features of each scale. The action representation at each scale is the L_1-normalized frequency of the occurrences of the features of that specific scale over the corresponding dictionary of visual words. The multiple scale-specific action representations are then fused either by their concatenation in one long representation vector $\mathbf{x}_C$ (Section 3.2) or by keeping them decoupled $\mathbf{x}_D$ (Section 3.3) until the matching stage.

4.4. Classification method

For action classification, we use a nonlinear SVM with the χ^2 distance metric ϕ for the matching of the action representation using the LibSVM toolbox [31].

$$\phi(\mathbf{x}^1, \mathbf{x}^2) = \frac{1}{2} \sum_{m=1}^{d} \frac{(\mathbf{x}_m^1 - \mathbf{x}_m^2)^2}{\mathbf{x}_m^1 + \mathbf{x}_m^2} \tag{1}$$

in which $\mathbf{x}^1$ and $\mathbf{x}^2$ are $d-$dimensional vectors which represent two different action representations based on d action primitives in the dictionary.

In the next section, we perform separability tests to evaluate the discrimination power of different types of action representations and motivate the use of multiple scale-specific representations over a single action representation. In Section 6, we perform classification testing to validate the classification accuracy of different representations on several benchmark datasets.

5. Separability test for action representation comparison

A separability test is performed to compare a single and multi scale-specific (Fig. 1(b) and 1(c)) representations. We thus use a discriminant criterion in which a good representation is obtained from a dictionary by which the samples from the same class will have very similar representations resulting in a low intra-class distance. In addition, the representation of samples

(a) Salient features- UCF sports

(b) Salient features- KTH

(c) Salient features- UCF Youtube

(d) Salient features- HOHA2

Figure 2: 2D projection of multi-scale asymmetric motion features on sample frames of (a) "diving" action from the UCF sports dataset [26] (b) "running" action from the KTH dataset [25], (c) 'play tennis' action from the UCF Youtube dataset [26], and (d) 'stand up' action from the HOHA [28] and HOHA2 [11] datasets. In (a), despite the movement of the camera, most of the motion features are detected from the athlete. In all video samples, the features are mainly from the moving body limbs and there are few false positives from the background.

from different classes should be very dissimilar in a sense that there is a high inter-class distance.

$$R = \frac{\sum_{k=1}^{N} \sum_{l \neq k}^{N} \phi^{kl}}{\sum_{k=1}^{N} \phi^{k}} \tag{2}$$

To implement the discriminant criteria, assume ϕ^{kl} denotes the average distances between the representations of the video samples from class k and l and ϕ^{k} denotes the average within-class distance of the representations of the video samples from class k. The discrimination score R (2) of a representation can thus be defined as the division of the sum of between-class distances for all N samples over the sum of within-class distances. When two representations are compared, the one with higher discrimination score is considered more discriminant as it has higher inter-class distances and lower intra-class distances. Higher score promises better separability of classes during classification.

5.1. Research question 1: separability of representations across scales

To evaluate the performance of representations across different spatio-temporal scales (research question 1), Fig. 3 shows the ratio of the discrimination scores between action representations at different scales (R_i/R_j) on the KTH and the UCF sports datasets. In Fig. 3(a), for example, the first row shows that the action representation of the finest scale $\mathbf{x}_1$ has higher discriminant score than any other scale (i.e., $R_1 > R_j$, $\forall j > 1$), and hence, it is the most discriminant representation. Overall, the main observation is that a finer-scale representation is more discriminant than a coarser-scale representation.

One could want to use just the finest scale representation as it is the most discriminant representation, but there are two factors that motivate the use of all scales. First, the distance from the camera and the speed of performance of actions might vary from one sample to another among the training/testing samples. More specifically, this spatial and/or temporal variation might require a representation at a different scale to be best matched with the finest scale of the training samples. Second, the action representation at each scale is constructed using a set of different dictionaries and different features. This diversity might be complementary for action representations and consequently, using just one representation might not necessarily be the best choice.

5.2. Research question 2: separability of single and concatenated representations

To evaluate the performance of encoding multi-scale spatio-temporal salient features using multiple scale-specific dictionaries versus a single non-scale-specific dictionary (research question 2), Fig. 4 shows the ratio of the discrimination scores of the concatenated representation $\mathbf{x}_C$ over the single action representation $\mathbf{x}_S$ for different actions from the KTH dataset and UCF sports dataset. For the KTH dataset, the $\mathbf{x}_C$ is always more discriminant than the $\mathbf{x}_S$ as the ratio is always higher than unit. For the UCF sports dataset, the $\mathbf{x}_C$ is most of the time more discriminant than the $\mathbf{x}_S$. The only exception is in discrimination of "walk" from "dive", "kick", and "run". The overall observation is that the concatenated representation $\mathbf{x}_C$ is more discriminant than a single action representation $\mathbf{x}_S$. The overall ratio for the KTH dataset is $R_C/R_S = 1.032$ and that for the UCF sports is $R_C/R_S = 1.02$ which shows the higher discrimination power of $\mathbf{x}_C$.

6. Action Classification Testing

This section presents the experimental results comparing the classification accuracy using the standard single representation and the introduced concentrated and decoupled representations in a common recognition framework on five benchmark dataset. We then provide a comparison with the state-of-the-art methods.

6.1. Research question 3: classification comparing single and new representations

Table 1 shows the average classification accuracy of single and two multi scale-specific representations using five benchmark datasets. The main observation is that both concatenated and decoupled representations provide higher classification accuracy than a single action representation. Also, the decoupled representation performs better than the concatenated fusion approach. This is due to incorporating the order of importance of representations in the matching and avoiding the possible curse of dimensionality problem of a concatenated representation.

Note that both HOHA and HOHA2 datasets have samples with multiple labels and hence, the action recognition in these datasets is a multi-label classification problem. To be consistent with the literature [11, 28, 32], we used the the one-against-all approach for performance evaluation and report the mean average precision as the classification accuracy.

6.2. Classification comparing new and existing methods

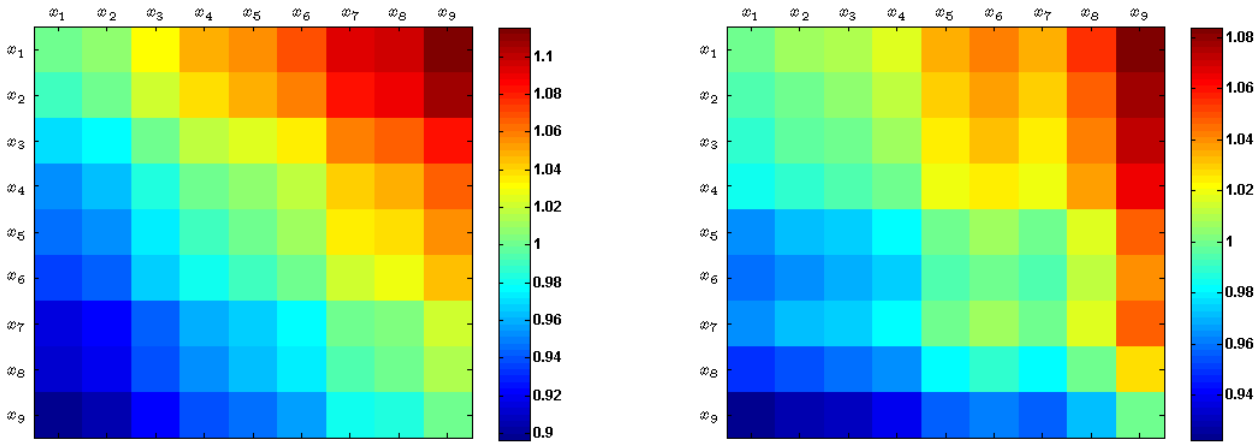
Even though the pipeline for the BOW framework is standard, there are several parameters in the implementation such as the SVM's similarity kernel, the dictionary length, and the features' dimensionality reduction which make direct comparison of different publications difficult. However, to show where our results stand in comparison with other published results, we provide Table 2 to compare the classification rate of different published methods which at least use the same training/testing protocols.

As Table 2 shows, the decoupled action representation with a single SVM classifier provides the highest classification accuracy in all benchmark human action recognition datasets, except the UCF youtube dataset. More specifically, the improvement on the choreographed KTH dataset is slightly better (about 1%) than the state-of-the-art methods, but our improvement is significant (about 5%) on realistic datasets of UCF sports, HOHA, and HOHA2 which are collected from youtube and Hollywood movies. Compared to the computationally expensive combined dense sampling and dense trajectory (DSDT) method (Wang et al. [33]), our approach performs about 0.5% less accurate which is not a big disadvantage as we used just a sparse set of salient features without any requirement for tracking. This shows that an effective discriminative action representation such as decoupled representation performs better than a computationally expensive method such as DSDT method [33].

When designing based on our improved action representations, the curse of dimensionality problem of the concatenated representation might be sever with increase in the number of scales and the length of the dictionaries. In such a scenario, the matching process of the decoupled approach might take longer as well. Based on the fact the spatial scale-invariant features are useful for object representation [29], but the temporal scale-invariance of the features might not be useful for motion content encoding [3], as a future research direction, one might try to investigate whether one could use scale-specific action representations of just temporal scales instead of each spatio-temporal scale. That is, the local salient features are spatially scale-invariant, but not temporally, and also the dictionaries are computed just for each temporal scales. As the number of temporal scales is typically fewer than the number of spatial scales, this policy reduces the number of representations which is very helpful to avoid or reduce the curse of dimensionality problem of the concatenated approach and the time complexity of the decoupled approach.

7. Conclusion

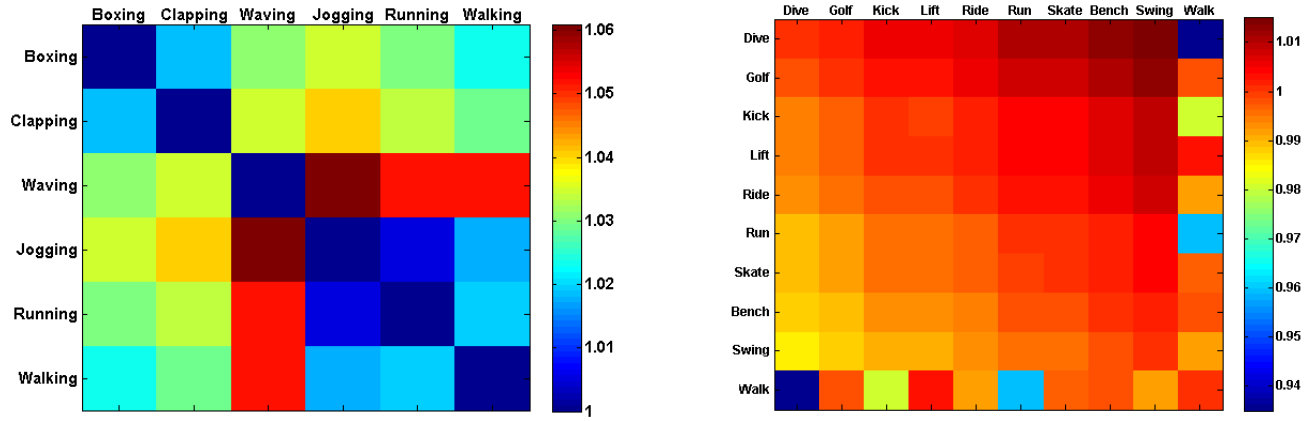
The standard BOW framework for human action recognition constructs a single dictionary of action primitives and hence,



(a) Separability test across scales- KTH

(b) Separability test across scales- UCF

Figure 3: The average ratio of the discriminant scores between action representations $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_9\}$ on (a) the KTH dataset and (b) the UCF sports dataset. Note that the matrices are reciprocal. To read this figure, each element of the matrix shows the ratio of the discriminant score of the representations with the corresponding row and columns index (e.g., the value at first row and fifth column shows R_1/R_5). Overall, a finer-scale representation is more discriminant than a coarser-scale representation. The plots are best viewed in color.



(a) Representations comparison- KTH

(b) Representations comparison- UCF sports

Figure 4: The average ratio of the discriminant scores of the concatenated multiple action representations $\mathbf{x}_C$ over the single action representation $\mathbf{x}_S$ on different actions from the (a) KTH dataset and (b) UCF sports dataset. As can be seen in these symmetric matrices, the $\mathbf{x}_C$ is always more discriminant than the $\mathbf{x}_S$ in the KTH dataset and most of the tim in the UCF sports dataset (the only exception is the separability of "walk" from "dive", "kick", and "run"). The plots are best viewed in color.

Table 1: Average classification accuracy on different datasets using the features detected by asymmetric motion features [4] and encoded by different action representation methods. The accuracy variation is in order of 0.01 and is not reported here. Note that both concatenated and decoupled action representations provide higher classification accuracy than a single action representation, regardless of the datasets. More specifically, the decoupled representation performs the best in all datasets.

Action representation	KTH	UCF sports	UCF youtube	HOHA	HOHA2
single	92.3 %	91.5 %	80.2 %	49.3 %	59.5 %
concatenated	94.1 %	92.3 %	81.5 %	52.3 %	61.3 %
decoupled	**95.3** %	**93.7** %	**83.7** %	**53.5** %	**63.2** %

Table 2: Comparison of different published methods for the human action classification on different benchmark datasets. Note that our method, the decoupled representation of asymmetric motion features with a SVM classifier, provides the highest classification accuracy in all datasets, except on UCF youtube in which the computationally expensive combined dense sampling and dense trajectory [33] performs slightly better than our efficient method which use a set of sparse salient features with no tracking component.

Method	*KTH*	*UCF sports*	*UCF Youtube*	*HOHA*	*HOHA2*
Wang et al. [33]	94.2 %	88.2 %	**84.2** %	-	58.3 %
Le et al. [34]	93.9 %	86.5 %	75.8 %	-	53.3 %
Wang et al. [9]	92.1 %	85.6 %	71.2 %	-	50.9 %
Rapantzikos et al. [35]	88.3 %	-	-	33.6 %	-
Laptev et al. [28]	-	-	-	38.4 %	-
Sun et al. [32]	-	-	-	47.1 %	-
Zhang et al. [36]	-	-	-	30.5 %	-
decoupled representation	**95.3** %	**93.7** %	83.7 %	**53.5** %	**63.2** %

a single action representation. We argued that a single representation cannot fully exploit the multiresolution characteristics of the different motion patterns. We therefore proposed the concept of learning multiple dictionaries of action primitives and consequently, multiple scale-specific action representations with two different fusion approaches of concatenation and decoupled. Having higher discrimination score, the scale-specific representations show better separability of different actions compared to standard single action representation. In a common recognition framework using multi-scale asymmetric motion features, both the decoupled and concatenated representations have shown superior improvements over the single representation on several benchmark human action recognition datasets including the KTH, UCF sports, UCF youtube, HOHA, and HOHA2 datasets. More specifically, the decoupled action representation with a single SVM classifier has shown about 5% improvement over the state-of-the-art methods on most of realistic datasets from youtube and Hollywood movies.

Acknowledgment

The authors would like to thank both GEOIDE (Geomatics for Informed Decisions), supported by the Natural Science and Engineering Research Council (NSERC) of Canada, and the Ontario Centres of Excellence (OCE) for financial support of this project.

[1] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal filters, IEEE International Workshop VS-PETS, Beijing, China (2005) 65–72.

[2] I. Laptev, B. Caputo, C. Schuldt, T. Lindeberg, Local velocity-adapted motion events for spatio-temporal recognition, Computer Vision and Image Understanding (2007) 207–229.

[3] G. Willems, T. Tuytelaars, , L. V. Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, European Conference on Computer Vision, Marseille, France (2008) 650–663.

[4] A. H. Shabani, D. A. Clausi, J. S. Zelek, Improved spatio-temporal salient feature detection for action recognition, British Machine Vision Conference, Dundee, UK (2011) –.

[5] R. Poppe, A survey on vision-based human action recognition, Image and Vision Computing 28 (2010) 976–990.

[6] W. Hu, N. Xie, L. Li, X. Zeng, S. Maybank, A survey on visual content-based video indexing and retrieval, IEEE Transactions on Systems, Man, and Cybernetics 41 (6) (2011) 797–819.

[7] M. Rodriguez, *CRAM*: Compact representation of actions in movies, IEEE Conference on Computer Vision and Pattern Recognition, San Fransico (2010) 3328–3335.

[8] J. Aggarwal, M. Ryoo, Human activity analysis: A review 43:16:116:43, 2011, ACM Computing Surveys (2011) 1–47.

[9] H. Wang, M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, British Machine Vision Conference, London, UK (2009) –.

[10] A. H. Shabani, J. Zelek, D. Clausi, Human action recognition using salient opponent-based motion features, IEEE Canadian Conference on Computer and Robot Vision, Ottawa, Canada (2010) 362–369.

[11] M. Marszalek, I. Laptev, C. Schmid, Actions in context, IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida (2009) 2929–2936.

[12] C. Bohm, S. Berchtold, D. Keim, Searching inhigh-dimensional spaces: Index structures for improving the performance of multimedia databases, ACM Computing Surveys 33 (3) (2001) 322–373.

[13] E. Hadjidemetriou, M. Grossberg, S. Nayar, Multiresolution histograms and their use for recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence 26 (2004) 831–847.

[14] F. B. N. Zouba, B. Boulay, M. Thonnata, Monitoring activities of daily living of elderly based on 3d key human postures, Int. Cognitive Vision Workshop (2008) 37–50.

[15] I. Laptev, On space-time interest points, International Journal of Computer Vision 64 (2005) 107–123.

[16] A. H. Shabani, D. A. Clausi, J. S. Zelek, Evaluation of local spatio-temporal salient features detectors for human action recognition, IEEE Canadian Conference on Computer and Robot Vision, Toronto, Canada (2012) 468–475.

[17] A. Shabani, D. Clausi, J. Zelek, Robust local video event detection for human action recognition, Neural Information Processing (NIPS), Machine learning for Assistive Technology Workshop, Whistler, BC (2010) –.

[18] T. Oommen, D. Misra, N. Twarakavi, A. Prakash, B. Sahoo, S. Bandopadhyay, An objective analysis of support vector machine based classification for remote sensing, Mathematical Geosciences 40 (2008) –.

[19] R. Bellman, Dynamic programming, Princeton University Press, 1957.

[20] N. Cristianini, J. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.

[21] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: A review, IEEE Trans. on Pattern Analysis and Machine Intelligence 22 (1) (2000) 4–37.

[22] N. Zheng, J. Xue, Statistical Learning and Pattern Analysis for Image and Video Processing, Springer, 2009.

[23] J. Weickert, A review of nonlinear diffusion filtering, International Conference on Scale-space theory in computer vision, Utrecht, Netherlands (1997) 3–28.

[24] H. Shabani, D. Clausi, J. Zelek, Towards a robust spatio-temporal interest point detection for human action recognition, IEEE Canadian Conference on Computer and Robot Vision, Kelowna, BC (2009) 237–243.

[25] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, IEEE International Conference on Pattern Recognition, Cambridge, UK 3 (2004) 32–36.

[26] M. Rodriguez, J. Ahmed, M. Shah, Action mach: a spatio-temporal maximum average correlation height filter for action recognition, IEEE Conference on Computer Vision and Pattern Recognition, Alaska (2008) 1–8.

[27] M. S. J. Liu, J. Luo, Recognizing realistic actions from videos "in the

wild", IEEE Conference of Computer and Pattern recognition, Miami, FL (2009) 1996–2003.

[28] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska (2008) 1–8.

[29] D. G. Lowe, Distinctive image features from scale-invariant key points, International Journal of Computer Vision 60 (2004) 91–110.

[30] P. Scovanner, S. Ali, M. Shah, A 3-*D*imensional *SIFT* descriptor and its application to action recognition, ACM Multimedia, Augsburg, Germany (2007) 357–360, `http://www.cs.ucf.edu/~pscovann/`.

[31] C. Chang, C. Lin, LIBSVM: a library for support vector machines, `http://www.csie.ntu.edu.tw/~cjlin/libsvm` (2001).

[32] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL (2009) 2004 – 2011.

[33] H. Wang, A. Klaser, C. Schmid, C. Liu, Action recognition by dense trajectories, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring (2011) 3169–3176.

[34] Q. Le, W. Zou, S. Yeung, A. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring (2011) 3361–3368.

[35] K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatiotemporal feature points for action recognition, IEEE Conference on Computer Vision and Pattern Recognition (2009) 1454–1461.

[36] T. Zhang, S. Liu, C. Xu, H. Lu, Boosted multi-class semi-supervised learning for human action recognition, Pattern Recognition 44 (2011) 2334–2342.