HUMAN ACTION RECOGNITION USING ASSOCIATED DEPTH AND SKELETON INFORMATION

Nick C. Tang[†] *Yen-Yu Lin*[⋆] *Ju-Hsuan Hua*[⋆] *Ming-Fang Weng*[⋆†] *Hong-Yuan Mark Liao*[†]

[†] Institute of Information Science, Academia Sinica, Taiwan
[⋆]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[⋆†]Smart Network System Institute, Institute for Information Industry, Taiwan

ABSTRACT

The recent advances in imaging devices have opened the opportunity of better solving computer vision tasks. The next-generation cameras, such as the depth or binocular cameras, capture diverse information, and complement the conventional 2D RGB cameras. Thus, investigating the yielded multi-modal images generally facilitates the accomplishment of related applications. However, the limitations of these devices, such as short effective distances, expensive costs, or long response time, degrade their applicability in practical use. Addressing this problem in this work, we aim at action recognition in RGB videos with the aid of Kinect. We improve recognition accuracy by leveraging information derived from an offline collected database, in which not only the RGB but also the depth and skeleton images of actions are available. Our approach adapts the inter-database variations, and enables the sharing of visual knowledge across different image modalities. Each action instance for recognition in RGB representation is then augmented with the borrowed depth and skeleton features.

Index Terms— Action recognition, Depth Association, Skeleton Association

1. INTRODUCTION

Most computer vision applications are established on image/video content analysis techniques, which are highly adapted to the available imaging devices. We are aware of the recent advances in imaging devices, such as the RGBD camera Microsoft Kinect[1]. The multi-modal images they provide give rich and diverse information. Thus, there has been a strong demand for content analysis techniques that leverage these cameras to better solve increasingly complex vision tasks, and even to initiate new applications. However, despite the great potential, these cameras have their restrictions. For instance, Kinect is with a short range of effective distance from 1.2 to 3.5 meters, and these cameras are relatively expensive. The restrictions hinder the applicability of these cameras in practical use. We address this issue in this work, and consider the scenario of solving a vision task with one online accessible image modality as well as a multi-modal dataset collected offline. We propose an approach that can borrow information from the extra dataset and facilitate the task of recognizing actions recorded by a 2D RGB camera.

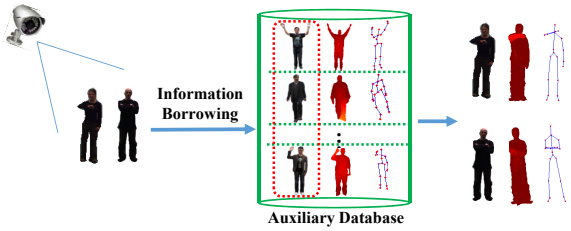

Fig. 1. The proposed framework.

As shown in Fig. 1, our main contribution in this work is to provide an effective way of utilizing new types of cameras, and better solve complex vision applications even when most of these cameras are not available online. The proposed approach is evaluated on a multi-view benchmark dataset which is captured by several RGB cameras. By using the same auxiliary database, our approach results in remarkable accuracy improvement in each dataset. Furthermore, our approach is developed in a general manner, and can be applied to vision applications in which multi-modal images are helpful.

2. RELATED WORK

Being one of the most important components in video understanding, action recognition is essential to widespread applications, such as surveillance. As indicated in [1], one fundamental difficulty of action recognition is the *large intra-class variations*. These variations can result from both intrinsic and extrinsic factors, including posture differences among subjects, clutter background, mutual or self occlusions. To address this problem, one of the current research trends in action recognition is to model the relationships among local features. For example, Matikainen et. al. [2] specify geometrical displacements between local features by generating

This work is supported in part by grants NSC 100-2221-E-001-013-MY3 and NSC 102-2221-E-001-025

[1]http://en.wikipedia.org/wiki/Kinect

a frequency lookup table. Besides, graphical models, such as *factorial conditional random fields* in [3] or *hidden Markov model* in [4], are applied to formulate the spatio-temporal correlation of local evidence. However, the afore-mentioned methods recognize actions based on 2D RGB images/videos. Restricted by information available, it is still very challenging to deal with intra-class variations caused by different camera perspectives or partial occlusions.

Owing to the recent advances in sensor technology, it is feasible to capture color as well as depth information of an action video in real time by RGBD cameras, *e.g.*, Kinect. Research efforts, such as [5–7], have demonstrated that depth maps of actions afford informative and invariant clues to build robust action recognition or pose estimation systems. Researches, *e.g.*, [8, 9], on 3D skeleton representation and correction may open the opportunity of resolving multi-view action recognition. The introduction of depth and skeleton information indeed benefits action recognition. However the short ranges of effective distances often make RGBD cameras inapplicable in real world applications, such as surveillance.

3. PROBLEM STATEMENT

We focus on recognizing actions of C classes. Suppose that we are given a training set, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{1, 2, ..., C\}$ are the RGB feature representation and the class label of the ith action, respectively. To borrow information across modalities, an auxiliary dataset, $A = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{d}}_i, \tilde{\mathbf{s}}_i)\}_{i=1}^{M}$, taken by Kinect is provided jointly, where $\tilde{\mathbf{x}}_i \in \mathcal{X}$, $\tilde{\mathbf{d}}_i \in \mathcal{D}$, and $\tilde{\mathbf{s}}_i \in \mathcal{S}$ are the RGB, depth, and skeleton feature representations of the ith instance, respectively. Note that auxiliary dataset A is unlabeled, and we use *tildes* to mark its data. With D and A, we aim to derive a good classifier for predicting test data that are similarly distributed to D. The auxiliary dataset A is collected to cover the action classes of interest in advance, *i.e.*, $\mathcal{Y}$ in this case. Building A beforehand is reasonable since we often focus on detecting some predefined types of actions in most action recognition applications. However, it is not necessary that the classes of actions in D and in A are the same. In addition, D and A can be established in different ways, so large inter-database variations are induced.

4. THE PROPOSED APPROACH

Our approach is composed of three stages, which are described as follows.

4.1. Cross-dataset correspondences

The goal of this stage is to correlate D and A, the two independently collected datasets, by exploring their common image modality, RGB. Specifically, we associate each $\mathbf{x}_i$ in D with a plausible $\tilde{\mathbf{x}}_{\pi_i}$ in A. A naïve way is the nearest neighbor search. However, it ignores the inter-database variations, and may result in sub-optimal performance. To address this issue, we exploit the data labels in D, and incorporate discriminant analysis to guide the construction of cross-dataset correspondences. We cast the task of cross-dataset correspondences as a labeling problem over Markov random fields (MRFs), in which the mutual verification among correspondences is activated. Hence, the borrowed multi-modal features are more discriminative.

In the construction of MRFs model with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each $\tilde{\mathbf{x}}_i$ in A corresponds to a *state*, while each $\mathbf{x}_i$ in D is associated with a variable node v_i, which takes a value from the state set $\mathcal{L} = \{1, 2, ..., M\}$. In this way, v_i determines the correspondence of $\mathbf{x}_i$ in A, *i.e.*$\{\mathbf{x}_i \in D, \tilde{\mathbf{x}}_{v_i} \in A\}$. An undirected edge $e = (v_i, v_j)$ is added into $\mathcal{E}$ if $\mathbf{x}_j$ is one of the ℓ nearest neighbors of $\mathbf{x}_i$. Hence, $|\mathcal{V}| = N$ and $\ell N/2 \leq |\mathcal{E}| \leq \ell N$. MRFs model the probability distribution over each possible labeling $V = [v_1 \cdots v_N] \in \mathcal{L}^N$ in form of

$$P(V) = \frac{1}{Z} \exp\left(-E(V)\right), \tag{1}$$

where *partition function* Z for normalization is defined as

$$Z = \sum_{V' \in \mathcal{L}^N} \exp\left(-E(V')\right). \tag{2}$$

In this work, we consider the following *energy function*:

$$E(V) = \sum_{v_i \in \mathcal{V}} \psi(v_i) + \sum_{(v_i, v_j) \in \mathcal{E}} \varphi(v_i, v_j), \tag{3}$$

where the *unary function* ψ and the *pairwise function* φ are defined as

$$\psi(v_i) = \begin{cases} \|\mathbf{x}_i - \tilde{\mathbf{x}}_{v_i}\|, & \text{if } \tilde{\mathbf{x}}_{v_i} \in k\text{NNs of } \mathbf{x}_i \text{ in } A, \\ \infty, & \text{otherwise,} \end{cases} \tag{4}$$

$$\varphi(v_i, v_j) = \begin{cases} \lambda_1 \|\tilde{\mathbf{x}}_{v_i} - \tilde{\mathbf{x}}_{v_j}\|, & \text{if } y_i = y_j, \\ -\lambda_2 \|\tilde{\mathbf{x}}_{v_i} - \tilde{\mathbf{x}}_{v_j}\|, & \text{otherwise,} \end{cases} \tag{5}$$

where kNNs denote the k nearest neighbors. k, λ_1, and λ_2 are three positive constants. They as well as ℓ are determined by cross validation. Note that only the labels of data in D are used. The auxiliary database A is assumed to be unlabeled.

The unary function in (4) ensures the compatibility of each correspondence. The pairwise function in (5) enforces class-consistent labeling. That is, $\tilde{\mathbf{x}}_{\pi_i}$ and $\tilde{\mathbf{x}}_{\pi_j}$ should be similar to each other if and only if $\mathbf{x}_i$ and $\mathbf{x}_j$ are of the same class. After applying *graph cut* [10] to minimizing (3), the most plausible configuration V is obtained. It follows that the N cross-dataset correspondences, $\{(\mathbf{x}_i, \tilde{\mathbf{x}}_{\pi_i})\}_{i=1}^{N}$, are established with $\pi_i \leftarrow v_i$.

4.2. Cross-modal feature association

In the stage, we aim to augment each training action in D and each testing action with additional depth and skeleton

features. Based upon the one-to-one modal mapping in A, the correspondences $\{(\mathbf{x}_i, \tilde{\mathbf{x}}_{\pi_i})\}_{i=1}^N$ established above can be propagated across image modalities, *i.e.*, $\{(\mathbf{x}_i, \tilde{\mathbf{d}}_{\pi_i})\}_{i=1}^N$ and $\{(\mathbf{x}_i, \tilde{\mathbf{s}}_{\pi_i})\}_{i=1}^N$. Yet, these correspondences are valid only for training data in D, and are not available for new testing data. To overcome this problem, we adopt *kernel canonical correlation analysis* (KCCA) to correlate data of two different domains, RGB $\mathcal{X}$ and skeleton $\mathcal{S}$, via $\{(\mathbf{x}_i \in \mathcal{X}, \tilde{\mathbf{s}}_{\pi_i} \in \mathcal{S})\}_{i=1}^N$.

Let $\phi : \mathcal{X} \to \mathcal{F}_x$ denote the feature map, which transforms data from domain $\mathcal{X}$ to space $\mathcal{F}_x$. Similarly, we have $\tilde{\phi} : \mathcal{S} \to \mathcal{F}_s$. Via ϕ and $\tilde{\phi}$, data of the two domains are mapped to high-dimensional Hilbert spaces, *i.e.*,

$$\mathbf{x}_i \mapsto \phi(\mathbf{x}_i) \quad \text{and} \quad \tilde{\mathbf{s}}_i \mapsto \tilde{\phi}(\tilde{\mathbf{s}}_i), \text{ for } i = 1, 2, ..., N. \tag{6}$$

KCCA seeks a pair of projections $(\mathbf{u}, \mathbf{v})$, and uncovers a common space, in which the correlation between projected data $\{\mathbf{u}^\top \phi(\mathbf{x}_i)\}$ and $\{\mathbf{v}^\top \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_i})\}$ is maximized. It has been proven in [11] that the projections lie in the span of data, *i.e.*,

$$\mathbf{u} = \sum_{i=1}^N \alpha_i \phi(\mathbf{x}_i) \quad \text{and} \quad \mathbf{v} = \sum_{i=1}^N \beta_i \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_i}). \tag{7}$$

In KCCA, the optimal projections $(\mathbf{u}^*, \mathbf{v}^*)$, parameterized by $(\boldsymbol{\alpha}^* = [\alpha_1^* \cdots \alpha_N^*]^\top, \boldsymbol{\beta}^* = [\beta_1^* \cdots \beta_N^*]^\top)$, are given by

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^\top K_x K_s \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^\top K_x^2 \boldsymbol{\alpha} \cdot \boldsymbol{\beta}^\top K_s^2 \boldsymbol{\beta}}}, \tag{8}$$

$$\text{where } K_x = [\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)] \in \mathbb{R}^{N \times N}, \tag{9}$$

$$K_s = [\tilde{\phi}(\tilde{\mathbf{s}}_{\pi_i})^\top \tilde{\phi}(\tilde{\mathbf{s}}_{\pi_j})] \in \mathbb{R}^{N \times N}. \tag{10}$$

The optimal $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ in (8) is obtained by solving a generalized eigenvalue problem. Furthermore, the formulation of KCCA can be extended to uncover multidimensional projections, *i.e.*, $U = [\mathbf{u}_1 \cdots \mathbf{u}_p]$ and $V = [\mathbf{v}_1 \cdots \mathbf{v}_p]$.

With U and V, we first project all the data under skeleton representation in A, *i.e.*$\{V^\top \tilde{\mathbf{s}}_i\}_{i=1}^M$. For an input action $\mathbf{x}$, which is either a training or a testing sample, we project $\mathbf{x}$ via $U^\top \mathbf{x}$, and retrieve its k nearest skeleton samples. The borrowed skeleton feature $\mathbf{s}$ is generated by minimizing the square reconstruction error, and is a convex combination of the k retrieved samples. We tune k via cross validation. The optimal range of k is $1 \sim 5$ in most of our experiments.

The same procedure is repeated for correlating image modalities RGB $\mathcal{X}$ and depth map $\mathcal{D}$. Every action $\mathbf{x}$ is augmented with two additional features borrowed from A:

$$\mathbf{x} \mapsto (\mathbf{x}, \mathbf{d}, \mathbf{s}). \tag{11}$$

It follows that the *augmented dataset* is constructed, *i.e.*, $D' = \{(\mathbf{x}_i, \mathbf{d}_i, \mathbf{s}_i)\}_{i=1}^N$.

4.3. Recognition with augmented features

The training data for action recognition have been expanded from D to D'. Three image modalities of each action are available at the same time. Early fusion or late fusion [12] can be adopted for combining the three heterogeneous features to achieve better performance. We have implemented both the two fusion strategies, and describe them in the following.

Multiple Kernel Learning for Early Fusion. We compile an kernel matrix for actions in each of the three image modalities, and adopt *SimpleMKL* [13], one of the state-of-the-art MKL packages, to learn an SVM classifier with multiple kernels. In this way, heterogeneous features are fused in the domain of kernel matrices.

Top-level Logistic Regression for Late Fusion. We learn an SVM classifier with probability estimation for each image modality, and concatenate the outputs of all the SVM classifiers. A top level $L2$-regularized logistic regressor is derived to work on data in this representation. In this manner, features are combined in the classifier level.

On Predicting A Test Action. Given a test action $\mathbf{x}$, we first augment it with the borrowed depth and skeleton features via (11). Then, either early fusion or later fusion can be applied to making the prediction.

Testing with data we collected, the performances of early fusion and late fusion are quite similar. Multiple kernel learning is less efficient owing to jointly tuning hyperparameters in kernel functions. Thus, we choose late fusion, and report quantitative results by late fusion in all the experiments.

5. IMPLEMENTATION DETAILS

To extract robust RGB features for action videos, we preprocess each video used in the experiments as follows. First, we apply the video inpainting technique [14] to compute the background images from a collection of sample videos. Then, we take the acquired background images as the mask, and adopt a background subtraction algorithm [15] to segment out the foreground region in each video frame. Accordingly, we can precisely compute the space-time volume (STV) features from the region of interest without worrying about the clutter background. In our implementation, we scale a given action video to the resolution $48 \times 64 \times t$ where t is the number of frames in the video. The *3D-HOG* (histogram of oriented gradients) descriptor [16] is applied to extract features both in a space-time volume and its horizontal mirror for against reflection. In more detail, we use $16 \times 16 \times 16$ pixel blocks, each of which is further divided into $2 \times 2 \times 2$ cells. Five hundred prototypes are derived to build up the embedding space. It leads to a compact representation for actions in RGB videos. As for depth features, we use the *Spatio-temporal Local Binary Pattern* (STLBP) as the feature representation of depth maps. The STLBP is developed to model the variation of motion and appearance based on concatenated LBP histograms. As for skeleton features, we implement the *Fourier Temporal Pyramid* [9] to represent the temporal dynamics of each 3D joint of a human body.

Table 1. Recognition rates (%) of different methods on the i3DPost dataset.

Method		Ours	RGB	DEP	SKE	[17]
Accuracy (%)	0°	95.2	86.9	88.1	80.9	77.5
	45°	96.4	91.6	83.3	84.5	84.9
	$0^\circ \cup 45^\circ$	94.7	85.1	88.7	86.3	84.9

6. EXPERIMENTAL RESULTS

To test the effectiveness of our approach, we present the performance of our approach to action recognition and compare it with other state-of-the-art methods. In addition, the benchmark of action recognition, i3DPost [18], is adopted in performance evaluation. This dataset contains 96 high-resolution RGB video sequences of 12 action types performed by 8 actors. These actions were recorded by multiple cameras with 8 different viewpoints. These actions recorded by multiple cameras with 8 different viewpoints. Each of these cameras was arranged to have 45° difference with its direct neighbors so that a full 360° coverage can be achieved.

Since our method performs visual knowledge borrowing across distinct data modalities, we use Microsoft Kinect to build up an multi-modal action database which contains the RGB frames, the depth maps and the corresponding skeletons and will be served as the common auxiliary database in the experiments on the benchmark. The auxiliary dataset is composed of 40 distinct action classes. Total 10 actors were employed in the construction of the dataset.Each action was recorded by two cameras, respectively located with view angles of 0° and 45°. Besides, we mirrored each recorded action for against reflection.

6.1. Baselines

For performance analysis and comparison, we implemented the following five baselines, each of which is denoted below in bold and in abbreviation:

RGB: This baseline simply ignores the information from the auxiliary database. It extracts the RGB features, described in Section 5, for the actions in the target database and employs an SVM classifier to make the prediction.

DEP: This baseline is a degenerate variant of our approach. Recall that our approach augments each RGB action video with additional depth maps. This baseline discards the original RGB features and the borrowed skeleton structures. It simply works on the borrowed depth maps. The adopted features for depth maps here are those described in Section V.

SKE: This baseline is the same as **DEP**, except the used data features are changed from the borrowed depth maps to the borrowed skeleton structures.

6.2. Experiment Settings and Quantitative Results

To make a fair comparison, we adopt the setup, Leave-One-Actor-Out (LOAO) cross validation, which is also used in [17]. The recognition rates of our approach, the three baselines, and the state-of-the-art systems are reported in TABLE 1. We consider three different cases for performance evaluation in benchmark i3DPost, including two singe-view settings (single-view 0° and single-view 45° for abbreviation), and one multi-view setting (multi-view $0^\circ \cup 45^\circ$ for abbreviation). Thus, there are three sets of quantitative results shown in TABLE 1, one for each case. It is worth mentioning some interesting observations. The baseline Bor-DEP and Bor-SKE are comparable or even better than baseline RGB. This phenomenon indicates that depth maps and skeleton structures are discriminative for actions in i3DPost. Our approach can effectively borrow features across video modalities, and leverage both the original and the borrowed features to result in much better accuracy rates. The performance gains of our approach over baseline RGB are significant in all the three settings, *i.e.*8.3% (95.2%-86.9%) in single-view 0°, 4.8% (96.4%-91.6%) in single-view 45°, and 9.6% (94.7%-85.1%) in multi-view $0^\circ \cup 45^\circ$. With the aid of cross-modal feature augmentation, our approach also remarkably outperforms the state-of-the-art system [17].

7. CONCLUSION

The new types of imaging devices provide the opportunity of better solving increasingly complex computer vision tasks, but their respective limitations are currently hindering the practical applicability. In the work, we resolve this problem by proposing an approach that can borrow information from an offline collected database where multi-modal images taken by heterogeneous cameras are available. Promising experimental results demonstrate that our approach can effectively adapt the variation between different databases, transfer knowledge across image modalities, and lead to remarkable performance boosting. In addition, the proposed approach is developed to carry out cross-modal information borrowing in a general way. It can be applied to a set of applications where multiple image modalities are appreciated, such as gesture recognition, human pose estimation, scene understanding, content-based multimedia analysis and recommendation.

8. REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, 2010.

[2] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," in *Proc. Euro. Conf. Computer Vision*, 2010.

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

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

[5] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. Int'l Conf. Computer Vision*, 2011.

[6] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011.

[7] X. D. Yang, C. Y. Zhang, and Y. L. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM Conf. Multimedia*, 2012.

[8] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, and Z. Tu, "Exemplar-based human action pose correction and tagging," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[9] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[10] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2001.

[11] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, 1998.

[12] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM Conf. Multimedia*, 2005.

[13] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Machine Learning Research*, 2008.

[14] N. C. Tang, C. T. Hsu, C. W. Su, T. K. Shih, and H.-Y. M. Liao, "Video inpainting on digitized vintage films via maintaining spatiotemporal continuity," *IEEE Trans. Multimedia*, 2011.

[15] O. Barnich and M. V. Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Processing*, 2011.

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

[17] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Trans. Neural Networks and Learning Systems*, 2012.

[18] N. Gkalelis, H. S. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost multi-view and 3d human action/interaction database," in *Proc. Conf. Visual Media Production*, 2009, pp. 159–168.