EVALUATING POSE ESTIMATION'S CONTRIBUTION TO VIOLENCE DETECTION USING DEEP LEARNING

SIAM TAHSIN BHUIYAN[1], MAHMUDUL HAQUE[1], HALIMA KHATUN[1], RASHEDUR RAHMAN[1], SAJED IMTENANUL HAQUE[2], ASHRAFUL ISLAM[1], SAADIA BINTE ALAM[1]

[1]Center for Computational & Data Sciences, Independent University, Bangladesh, Dhaka 1229, Bangladesh
[2]Department of Computer Science and Engineering, Independent University, Bangladesh, Dhaka 1229, Bangladesh
E-MAIL: siamtbhuiyan@gmail.com, mahmud.eece@gmail.com, halimakhatunmy@gmail.com, rashed.riyadh14@gmail.com, imtenanul@iub.edu.bd, ashraful@iub.edu.bd, saadiabinte@iub.edu.bd

Abstract:

This study investigates how pose estimation can enhance violence detection using the RWF-2000 dataset and a ConvLSTM-based approach. Pose estimation was integrated into the method by employing MoveNet to extract skeletal key points from video frames, enabling the system to focus on movement patterns for more effective detection of violent actions. Two versions of the model were trained: one using the original dataset and another with a pose-estimated variant. The results demonstrate that the pose-estimation variant achieved higher accuracy, precision, recall, specificity, and F1 score compared to the base version. These improvements indicate that pose estimation enhances the model's ability to interpret movements by minimizing irrelevant background details and emphasizing critical motion patterns. This research highlights the potential of skeletal data to improve the reliability and accuracy of violence detection, supporting advancements in surveillance and action recognition systems.

Keywords:

Violence detection; Pose estimation; ConvLSTM; MoveNet; Action Recognition; RWF-2000 dataset; Skeletal keypoints; Deep learning.

1. Introduction

The detection of violence through image-based classification has become a focal point in Deep Learning (DL) research [1]. The widespread adoption of surveillance technologies, such as CCTV, Body-Worn Cameras, and Smartphone cameras has enabled the development of Automatic Violence Detection (AVD) systems. These systems are important for real-time surveillance in densely populated areas which are prone to violent incidents. It enhances law enforcement capabilities and ensures community safety. Additionally, the rapid spread of graphic content, such as videos containing violence can trigger harmful social behaviors, underscoring the urgent need for reliable AVD systems to maintain public order [2], [3].

Traditional violence detection methods relied on handcrafted features like Histograms of Oriented Gradients (HOG) and optical flow. They struggled to handle the complexities of human interactions and variable backgrounds often present in real-world videos [4]. These limitations encouraged the adoption of deep learning methods, with Convolutional Neural Networks (CNNs). They are becoming a preferred approach for extracting spatial features from images [5]. CNNs have demonstrated their effectiveness in activity recognition tasks by capturing rich spatial information from video frames [6]. However, CNN-only models fall short when it comes to modeling temporal dynamics essential for understanding violent behaviours. This gap has led to the exploration of architectures that integrate both spatial and temporal modeling.

To address the need for spatiotemporal analysis, researchers have proposed models like 3D CNNs, which extend convolutional operations to the temporal dimension. While 3D CNNs effectively learn spatiotemporal features, they are computationally expensive and can overfit when trained on limited datasets [7]. The Hierarchical Recurrent Neural Network (HRNN) for skeleton-based action recognition introduces a structure where human joint data is processed hierarchically to retain spatial hierarchies while modeling temporal sequences [8]. By grouping joints in a way that preserves their spatial relationships before applying recurrent layers to capture temporal dependencies, this approach efficiently interprets complex motion patterns. An alternative approach is the ConvLSTM model, which extends traditional LSTMs by introducing convolutional operations to process video data while preserving spatial information. ConvLSTM has shown ability to efficiently manage spatiotemporal sequences, making it well-suited for video analysis tasks [9], [10].

Despite these advancements, accurately detecting violent activities remains challenging under varying conditions, such as occlusions, low lighting, and background noise. These factors disrupt a model's ability to focus on essential human movements. By focusing on human skeletal features, pose estimation reduces background distractions and emphasizes movement-related features. Methods like OpenPose and MoveNet are particularly effective at tracking key body points, improving action recognition when appearance-based cues are insufficient [11], [12].

In violence detection, the integration of pose estimation has shown promising results. Studies such as [13] have used skeletal data extracted with OpenPose, demonstrating enhanced performance compared to models trained on raw video data. Similarly, a study highlighted that skeletal features can capture critical cues of violent movements, although challenges like managing multiple individuals and achieving real-time performance remain [14]. Comparative studies have shown that while pose-based models often outperform traditional models in crowded settings, they may struggle when skeletal information alone is inadequate for capturing complex interactions [15].

In this study, we investigate the impact of pose estimation on violence detection performance using the RWF-2000 dataset. The main contributions of this paper are:

1. Enhanced violence detection by integrating MoveNet-based pose estimation with ConvLSTM to improve spatio-temporal feature extraction.
2. Conducted a detailed comparison of models trained on original and pose-estimated RWF-2000 datasets, evaluating performance metrics such as accuracy, precision, recall, and F1 score.
3. Developed an efficient, scalable framework for surveillance and content moderation, addressing challenges like occlusions and background noise through pose-enhanced video analysis.

2. Datasets

In this study, we used the RWF-2000 dataset [16], which consists of 2000 short videos categorized into two classes: Violent and Non-Violent. Each short video is 5 seconds long, recorded at 30 frames per second, and sourced from YouTube. These short videos are derived from authentic surveillance footage, encompassing a diverse range of resolutions and lighting conditions. The dataset captures various forms of violence, including one-on-one Violents and crowd disturbances, across different settings such as indoor and outdoor environments. This diversity and realism make the RWF-2000 dataset an excellent choice for developing and evaluating a robust violence detection model.

3. Proposed Method

This study explores the transformative potential of pose estimation in enhancing violence detection performance using the RWF-2000 dataset. We develop a two-stage approach where in the original dataset undergoes pose estimation preprocessing, extracting skeletal body landmarks that capture detailed movement characteristics. Two identical ConvLSTM models are subsequently trained—one on the original dataset and another on the pose-estimated variant, enabling a rigorous comparative evaluation of detection capabilities. By maintaining a consistent architectural framework, we isolate the impact of pose estimation, providing insights into how skeletal representations can potentially improve the computational understanding of violent actions. This approach not only addresses the computational challenges in violence detection but also demonstrates the value of preprocessing techniques in enhancing deep learning model performance. The comparative analysis offers a comprehensive assessment of pose estimation's efficacy, potentially opening new avenues for intelligent video surveillance and action recognition systems. Figure 1 provides a comparative overview of two models demonstrating the method overview.

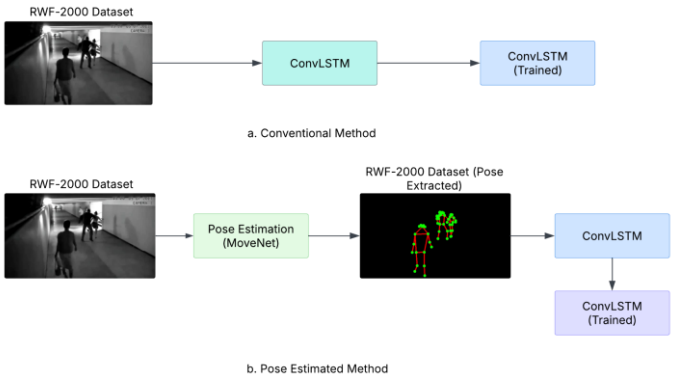

FIGURE 1. Method Overview

3.1. Dataset Preprocessing

In the initial step, the video files are divided into 24 frames, where each frame has a shape of (120, 120, 3). Each set of frames is paired with a label indicating whether the video shows a violent or a Non-violent scenario. The data is then shuffled to ensure randomization. The RWF-2000 dataset

consists of a total of 2000 videos, with an equal split between violent and non-violent categories. From this dataset, 1600 videos (800 violent and 800 Non-violent) are added to the training set, 200 videos (100 violent and 100 non-violent) to the validation set, and the remaining 200 videos (100 violent and 100 non-violent) to the test set. Finally, the data is batched with a batch size of 16, resulting in an 80% training, 10% validation, and 10% testing split of the overall dataset.

3.2. Pose Estimation

For pose estimation, we employed TensorFlow's pretrained "MoveNet" model, a state-of-the-art approach that identifies 17 key points on the human body. MoveNet operates as a bottom-up estimation model, utilizing heatmaps to precisely locate these key points. Its architecture consists of two main components: a feature extractor and prediction heads. The prediction head is loosely inspired by CenterNet [17], while the feature extractor is a modified version of MobileNet [18]. MoveNet offers two variants, Lightning and Thunder; we selected the multipose version of the Lightning variant to enable detection of multiple individuals within each frame. To create a pose-estimated version of the RWF-2000 dataset, we processed each video by iterating through its frames and applying the MoveNet model on each one. For each detected person in a frame, we annotated the key points and connected them accordingly. Once all persons in a frame were estimated, we removed the background, leaving only the skeletal representations. This process was repeated across the dataset, resulting in the Pose Estimated RWF-2000 dataset.

Following the pose estimation process, Figure 2 visually demonstrates the transformation of the RWF-2000 dataset through the application of the MoveNet model. The first column represents the original input frames, while the second column displays the pose-estimated outputs with the background retained. The final column illustrates the processed frames with background removal, leaving only skeletal representations of detected individuals.

If we assess its performance across various scenarios, the first row demonstrates the model's effectiveness in accurately identifying and connecting key points on a single individual, highlighting its precision under ideal conditions. The second row shows the model's performance in group settings, where closely positioned or overlapping individuals introduce moderate challenges, occasionally leading to minor inaccuracies in key point localization. The third row illustrates the model's limitations, where factors such as extreme poses and occlusions result in significant errors in key point estimation. These cases collectively highlight MoveNet's strengths and its areas for improvement.

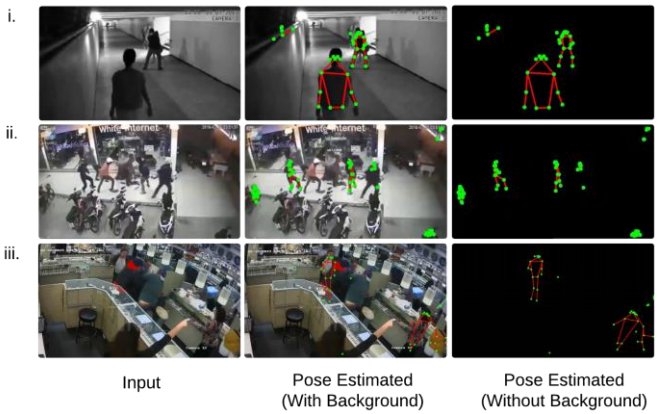

FIGURE 2. Pose Estimated Representation

3.3. Violence Detection

For the violence detection component, we selected the ConvLSTM architecture due to its superior ability to handle spatiotemporal data. ConvLSTM builds upon the foundation of Fully Connected LSTMs (FC-LSTM) by incorporating convolutional structures into both the input-to-state and state-to-state transitions [19]. As demonstrated in [17], stacking multiple ConvLSTM layers to form an encoding-forecasting framework enables the network to effectively manage general spatiotemporal sequence forecasting tasks. In contrast, FC-LSTMs operate on one-dimensional vectors, which limits their capacity to maintain spatial relationships inherent in image and video data, rendering traditional LSTMs unsuitable for such applications [20]. ConvLSTM offers a more efficient alternative by requiring fewer parameters while still preserving spatial hierarchies, making it particularly well-suited for processing images and videos. This balance of reduced complexity and maintained spatial integrity makes ConvLSTM an ideal choice for our violence detection model, allowing it to accurately capture and interpret the dynamic interactions present in visual data.

For the convolutional backbone of our ConvLSTM architecture, we selected the "ResNet50V2" model [21], a 50-layer variant of the widely used ResNetV2 architecture. Residual Networks, first introduced in [22], gained popularity for their ability to mitigate vanishing and exploding gradient issues when increasing network depth, thereby improving accuracy. Previously, increasing the layer count in deep networks often led to accuracy degradation due to the difficulty of back propagating gradients through multiple layers. Residual Blocks address this issue by introducing skip connections, allowing gradients to bypass certain layers and effectively preserving gradient flow [22]. In ResNetV2 [21], the architecture was refined by introducing batch

normalization before each weight layer, a modification that improved performance. Various ResNetV2 architectures are available, including ResNet50V2, ResNet101V2, and ResNet152V2. Among these, we selected ResNet50V2 for its balanced trade-off between complexity and performance. We utilized a pre-trained ResNet50V2 model, originally trained on the ImageNet dataset [23], as the convolutional feature extractor within the ConvLSTM framework. This integration enabled us to leverage ResNet50V2's robust feature extraction capabilities for spatiotemporal analysis in violence detection.

After initializing the sequential model, we started by adding the ResNet50V2 layer as the primary convolutional backbone. Next, we incorporate a TimeDistributed Flatten layer to handle sequential data across time steps, followed by a Bidirectional LSTM layer with 128 units to capture temporal dependencies in both forward and backward directions. To prevent overfitting, we included two Dropout layers with a dropout rate of 0.5. Between these, we add a Dense layer with 128 units and "ReLU" activation to enhance non-linear transformations. Finally, we complete the model with a single-unit Dense layer with "Sigmoid" activation, providing a binary output for classification.

3.4. Model Setting

Given that the model performs binary classification (Violent vs. Non-Violent), we compile it using the Binary Cross-Entropy loss function. This loss function quantifies the difference between the model's predicted probability distribution and the true distribution for each class. The Binary Cross-Entropy loss L for a single instance is expressed in equation 1.

$$L = -\frac{1}{N}\sum_{i=1}^{N} yi \times \log\big(p(yi)\big) + (i - yi) \times \log\big(1 - p(yi)\big) \quad (1)$$

In this context, yi represents the actual class label, where p(yi) is the predicted probability of the instance belonging to class 1 (Violent), and 1- p(yi) represents the probability of it belonging to class 0 (Non-Violent).

We used ReduceLearningRateOnPlateau with default settings to adjust the learning rate during training. To minimize Binary Cross-Entropy loss, we applied the Adam optimizer with an initial learning rate of 0.000001. The model was trained for 100 epochs with a batch size of 16 to ensure effective learning and stable performance.

3.5. Evaluation

To evaluate the model's performance, we compute key metrics including Accuracy, Precision, Recall, Specificity, and F1 Score. Each of these values is calculated using True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). In this context, True Positives (TP) represent instances of violent activity that are correctly classified as violent, while False Positives (FP) denote non-violent instances that are incorrectly classified as violent. True Negatives (TN) correspond to non-violent activities that are accurately identified as non-violent, and False Negatives (FN) refer to cases of violent activity that are mistakenly classified as non-violent.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$F1\ score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (6)$$

We compute Accuracy, Precision, Recall, Specificity, and F1 Score using equations 2 - 6 to evaluate the model's performance. Accuracy represents the proportion of correct predictions among all predictions, offering a general measure of model effectiveness. Precision indicates the accuracy of violent activity predictions by measuring the ratio of correctly predicted violent instances to all instances predicted as violent. Recall measures the model's ability to identify actual violent instances by calculating the ratio of correctly predicted violent cases to all true violent cases. Specificity provides insight into the model's accuracy in identifying non-violent cases, calculated as the ratio of correctly predicted non-violent instances to all actual non-violent instances. Finally, F1 Score, as the harmonic mean of Precision and Recall, provides a balanced measure of the performance, particularly useful in cases where class distributions are imbalanced. A high F1 Score indicates that the model maintains a strong balance between correctly identifying positive instances and minimizing misclassifications.

4. Result & Discussion

After training the models for 100 epochs, their performance was assessed using a confusion matrix, which quantified true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) counts. This evaluation on the test dataset offered valuable insights into the model's discriminative ability, highlighting its effectiveness in distinguishing violent from non-violent scenarios across various performance metrics.

The experimental results in Table I demonstrate that the pose-estimated model consistently outperforms the base model across all metrics, with a 14% increase in precision and specificity, reducing false positives and improving non-violent case detection. Recall improved by 2%, slightly enhancing the detection of violent instances, while the F1 score also increased, reflecting a better balance between precision and recall. These results validate the effectiveness of pose estimation in refining the model's discriminative ability and improving violence detection accuracy.

TABLE 1. Comparative Performance Evaluation

Dataset	Accuracy	Precision	Recall	Specificity	F1 Score
RWF-2000 (Base)	68.50%	65.81%	77%	60%	0.709
RWF-2000 (Pose Estimated)	**77.08%**	**79.23%**	**79%**	**74%**	**0.770**

The confusion matrix analysis from figure 3 shows incremental improvements in classification performance through pose estimation. Without pose estimation, the model

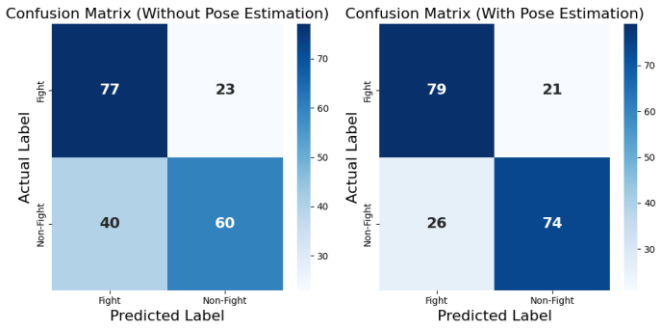

FIGURE 3. Confusion Matrix (Without Vs With Pose Estimation)

correctly classified 77 out of 100 violent cases and 60 out of 100 non-violent cases, revealing notable misclassification challenges. Conversely, the pose-estimated model demonstrated enhanced performance, accurately identifying 79 violent cases and 74 non-violent cases, representing a meaningful improvement in detecting both violent and non-violent scenarios. These performance gains come from pose estimation's ability to simplify video data to its most essential elements. By systematically eliminating background noise, variable lighting conditions, and extraneous visual information, the technique enables the model to concentrate exclusively on human skeletal movements—the most critical features for violence detection. MoveNet's sophisticated extraction of 17 key body points effectively strips away peripheral visual data, leaving only the core kinematic information essential for precise action recognition.

However, the study has certain limitations. Pose estimation errors (inaccurate or missing keypoints) led to incorrect motion representations, adversely affecting classification accuracy. Improving the accuracy of pose estimation could further enhance the model's overall detection performance. Additionally, pose estimation is particularly effective for human-to-human violence detection but may struggle when objects are involved, potentially reducing detection accuracy. While the observed improvements confirm our hypothesis, there is still room for further refinement in the model's performance. Future work will focus on exploring advanced architectures such as Vision Transformers (ViTs) or custom models to enhance accuracy. Developing specialized pose estimation models tailored for violence detection could reduce pose estimation errors, further improving detection reliability and overall performance.

5. Conclusion

This study underscores the significant impact of pose estimation on violence detection, utilizing the RWF-2000 dataset and a ConvLSTM architecture. By integrating pose-estimated skeletal data, the model demonstrated enhanced performance across all evaluation metrics, including accuracy, precision, recall, specificity, and F1 score. The observed improvements can be attributed to the effectiveness of pose estimation in enhancing action recognition by filtering out irrelevant background noise and emphasizing key human movement patterns. Furthermore, the study highlights the importance of applying appropriate preprocessing techniques in deep learning pipelines, which play a critical role in optimizing model performance. While these results validate the hypothesis regarding the benefits of pose estimation, there remains room for further refinement. Developing specialized pose estimation models specifically tailored for violence detection in videos could help reduce estimation errors and enhance overall detection reliability, paving the way for future advancements in computational video analysis methods.

Acknowledgements

This paper is partially supported by a grant from the Independent University, Bangladesh (IUB).

References

[1] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on

state-of-the-art violence detection techniques", IEEE Access, Vol. 7, pp. 107560-107575, 2019.
[2] S. Liu and T. Forss, "New classification models for detecting hate and violence web content", Proceeding of IC3K2015 Conference, Lisbon, Portugal, Vol. 1, pp. 487-495, November 2015.
[3] C. Gu, X. Wu, and S. Wang, "Violent video detection based on semantic correspondence", IEEE Access, Vol. 8, pp. 85958-85967, 2020.
[4] R. Poppe, "A survey on vision-based human action recognition", Image and Vision Computing, Vol. 28, No. 6, pp. 976-990, June 2010.
[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks", Proceeding of CVPR2014 Conference, Columbus, OH, USA, pp. 1725-1732, June 2014.
[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos", Proceeding of NeurIPS2014 Conference, Montreal, Canada, pp. 568-576, December 2014.
[7] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 1, pp. 221-231, January 2013.
[8] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton-based action recognition", Proceeding of CVPR2015 Conference, Boston, MA, USA, pp. 1110-1118, June 2015.
[9] X. Shi et al., "Convolutional LSTM network: A machine learning approach for precipitation nowcasting", Proceeding of NeurIPS2015 Conference, Montreal, Canada, pp. 802-810, December 2015.
[10] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs", Proceeding of ICML2015 Conference, Lille, France, pp. 843-852, July 2015.
[11] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks", Proceeding of CVPR2014 Conference, Columbus, OH, USA, pp. 1653-1660, June 2014.
[12] S. Girdhar, A. Ramanan, and J. Malik, "Attentional pooling for action recognition", Proceeding of ECCV2018 Conference, Munich, Germany, pp. 34-49, September 2018.
[13] Y. Zhang, "Violence detection using skeleton-based action recognition", Proceeding of ICIP2020 Conference, Abu Dhabi, UAE, pp. 3483-3487, October 2020.
[14] M. Ke, "Pose-based violence detection in surveillance videos", IEEE Access, Vol. 8, pp. 19327-19335, 2020.
[15] H. Wang et al., "Comparative analysis of pose estimation-based and appearance-based action recognition", Proceeding of ICCV2019 Conference, Seoul, South Korea, pp. 21-30, October 2019.
[16] M. Cheng, K. Cai, and M. Li, "RWF-2000: An Open Large Scale Video Database for Violence Detection", Proceeding of ICPR2020 Conference, Milan, Italy, pp. 4183-4190, January 2021.
[17] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as Points", arXiv:1904.07850 [cs], April 2019.
[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks", Proceeding of CVPR2018 Conference, Salt Lake City, UT, USA, pp. 4510-4520, June 2018.
[19] X. Shi, Z. Chen, H. Wang, and D.-Y. Yeung, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting", Proceeding of NeurIPS2015 Conference, Montreal, Canada, Vol. 28, December 2015.
[20] Soheil Vosta and Kin-Choong Yow, "A CNN-RNN combined structure for real-world violence detection in surveillance cameras", Applied Sciences, Vol. 12, No. 3, pp. 1021, 2022.
[21] Kaiming He et al., "Identity mappings in deep residual networks", Proceeding of ECCV2016 Conference, Amsterdam, Netherlands, Part IV, pp. 630-645, October 2016.
[22] Kaiming He et al., "Deep residual learning for image recognition", Proceeding of CVPR2016 Conference, Las Vegas, NV, USA, pp. 770-778, June 2016.
[23] Jia Deng et al., "ImageNet: A large-scale hierarchical image database", Proceeding of CVPR2009 Conference, Miami, FL, USA, pp. 248-255, June 2009