Annotate And Visualize A Video With Pose Recognition And Text Gloss Ssg

S. Selvi

Department of computer science and engineering, K. S. R College of Engineering, Tiruchengode.
rsselvi.ammu@gmail.com.

Abstract: Good interfaces should help us to decide what next? Hence, Interface design issues pertinent to this project include the integration of media, annotation, use of metaphors, standards and the goals of the multimedia system. Without control over facial appearance, screening perspective, illumination circumstances, occlusions, background, expression, hairstyles, in the video quality is exaggerated and image contains noise pixels and Face data in different angle cannot be recognized. So recognize an unimpeded pose and visualizing a video in multimedia environment with authoring is, still a thorny way. In this paper we present a dual course of action, at first annotate a video with excellent pretense extraction using Fuzzy Local Information C-Means (FLICM) Clustering Algorithm. This approach is used to fuzzy local gray level quantify resemblance, aiming to assure noise insensitiveness and image detail preservation. Subsequently we can generate sketches using sobel operator which finds the sharp intensity variation of an image and it obtains the edges of the objects contained on the image. At the jiffy phase we propose a text gloss scene structured graph to authorize a video.

Keywords: **Annotation, Authoring, Pose recognition, Sketches, Visualization**

I. INTRODUCTION

Multimedia interfaces are, “operates communally”. Interface is a interactive plummet between the first and transitional sense. In high-quality multimedia interface should integrate different media types without a glitch. Hence the interface must provide smooth transitions between media types. Multimedia recurrently uses the metaphor of human-to-human communication as part of the interface design. We intend a innovative way to swiftly and precisely envisage poses their individual questions sketch out their own tactics for deeper analysis and authorize a video with text gloss SSG. Annotation is a metadata which contains state, elucidation, presentational score fond to text, image or supplementary multimedia data. It refers to the explicit division of the unique data. Exertion on video annotation had started in 1992 & researchers are still working on shaping an efficient video annotation techniques in reverence of database & time required to annotate the video. Video annotation generally aim at the task of solitary or manifold concept labels to a target data set, where the assignment is often done autonomously without considering the inter-concept liaison. It aims to assign several semantic & visual features to the contents of video and adding the descriptors for the contents of video and background information about the video, and hence it facilitate semantic retrieval of videos from a hefty video database. Sketches are an expressive and discerning query medium and it can portrays the vital aspects. So we have implemented the notion of sketches, which are helpful to succinctly annotated with processes, concepts, and interrelationships, in addition to labels of features. The foremost aspects of a sketches annotated with concise is (1) recognize the features, (2) portray the processes that are stirring, and (3) exemplify the affairs between features and processes. Pose identification has been one of the most important areas of research in the meadow of computer vision and pattern recognition. Because it is the intricate to determine, whether two face images portray the same person or not. Our focal contribution is to treat pose evaluation as object recognition, using a narrative sketch representation which spatially concentrate joints of interest at low computational cost and high accuracy and our highly varied training dataset allows the classifier to estimate body parts invariant to pose, body shape, clothing, etc. The desired transformation is learned using MDS (Multi-Dimensional Scaling). It is a method that represents measurements of similarity (or dissimilarity) among pairs of objects as distances between points of a low-dimensional space. Side information of feature distances between images had they all been of the same resolution and pose is provided to assist in learning the transformation that maps the difficult pose estimation problem. The law of visibility is, the user should proficient to glimpse the deed that are unbolt to them at every point of occasion. It afford instant response about the action taken and acquire shrewd information about the sequence of action. Authoring context sensitive, interactive multimedia presentations is much more complex than authoring either purely audiovisual applications or text. Video authoring is a design process. It is preferred by users to hastily investigate, compare, and commune diverse design ideas with high-level semantic information in an early design process [1]. The function of multimedia authoring is that people converse message with each other using assorted media forms. Preceding toil on video authoring uses design primitives including captions, key frames, and videos. Captions, as well as text annotations, can provide valuable semantic information for understanding media. So we afford text gloss SSG. Ultimately, the video authoring can be achieved by integrating related video sources based on the visual layout structures. The rest of the manuscript is structured as follows. An indication of the related approaches is discussed in Section II. A pithy depiction of the sketch and text based annotation is provided in Section III. The minutiae of the pose recognition are provided in Section IV. Section V discusses the approach of scene structured graph for authoring a video. The conclusion are presented in Section VI.

II. RELATED APPROACH

Numerous approaches have been anticipated in the literature for managing individual or added factors like sketching a image, recognizing a pose, illumination and resolution which affect face recognition performance and

authoring a video. Video content annotation is gaining substance at present, for effortless salvage of requisite video from hefty quantity of available sources. At the early stages annotation of image/video content was prepared manually by user by navigating through various frames of videos. So at present different techniques are proposed. Automatic annotation is made by edifice replica based on low-level features for apiece of keyword in a terminology, e.g., the multiple Bernoulli relevance model in [7]. Hierarchical topic trajectory model (HTTM) is proposed for acquiring a dynamically changing topic model which represents relationship between video frames & associated text labels. A graph reinforcement method is used to resolve the involvement of a akin manuscript to the annotation goal[15]. The image can be sketched by an assortment of methods in offered technologies such as, laplacian pyramid,coherent line drawing algorithm, prewitt operator, sobel operator,etc. Pose recognition can be performed by Baker and Kanade they propose an algorithm to learn a prior on the spatial distribution of the image gradients for frontal images of faces. Bhatt presents a Local Binary Patterns[9],Prince *et al.* [8] propose a generative model for generating the scrutiny space from the identity space using an affine mapping and pose information. Baker and Kanade [10] propose an algorithm to learn a prior on the spatial distribution of the image gradients for frontal facial images. The underlying principle of multimedia authoring is that people converse message with every one using assorted media forms. Preceding exertion on video authoring [11] [12] uses design primitives with texts, captions, keyframes, and videos. Captions, as well as text annotations, can provide precious semantic information for indulgent media [13], [14]. The ultimate yield of the authoring process, we used MPEG-7 [6]

III. SKETCH AND TEXT-BASED GLOSSING

The idea of using sketching and gestures are, interacting with computers. There are many special sketch-based interfaces facilitate at present. Therefore, it is important to be able to discriminate and label them in some mode. Two of the most significant characteristics of sketch-based interfaces are the quantity of strokes the computer looks next to make an elucidation and the underlying ambiguity level. The ambiguity level refers to how tricky it is to interpret a sketch given the generality or specificity of the domain. In other words, a sketch-based interface can have a high ambiguity level if there are many possible interpretations the computer could find for any one particular sketch, and this often occurs with very general domains. Restricting the domain to be very limited in scope and can reduce the ambiguity level.

A. SKETCH TO POSE RECOGNITION USING SOBEL OPERATOR

Sketches and images can be viewed as two diverse modalities. So, it is necessary to pertain various transformations that can diminish the disparity between sketches and digital images. Since, a face sketch is principally an edge depiction of the tangible face in which prominent edges are tinted, edge preserving techniques can be used for this chore. In the proposed approach, sobel operator is used for the sketch-digital image pairs to conserve edges. Sobel operator performs a 2-D spatial gradient depth on an image and emphasizes regions of soaring spatial gradient that correspond to edges. Classically it is used to locate the ballpark gradient magnitude at apiece point in an input grayscale image. Compared to other edge operator, Sobel has two focal recompense:

1. As the prologue of the average factor, it has some smoothing result to the haphazard din of the image.
2. It is the differential of two rows or two columns, so the elements of the edge on both sides has been enhanced, so that the edge seems thick and bright. The difference between the original and sobel applied image is described in figure 1 and 2.The sketch generation in digital image algorithm is described as follows:

Masks can be applied discretely to the input image, to fabricate detach dimensions of the gradient module in apiece orientation as Gx and Gy. These can then be pooled mutually to locate the absolute magnitude of the gradient at apiece point and the orientation of that gradient. Although characteristically, an estimated magnitude is computed using:

$$|G| = |Gx| + |Gy|$$

PSEUDO-CODES FOR SOBEL EDGE DETECTION
Input: A taster Image.
Output**:** Detected Edges.
Step 1: Acknowledge the taster image.
Step 2: Concern mask Gx,Gy to the taster image.
Step 3: Apply Sobel edge detection algorithm and the gradient.
Step 4: Masks handling of Gx,Gy separately on the taster image.
Step 5: Outcome pooled to locate the absolute magnitude of the gradient.
Step 6: The absolute magnitude is the output edges.

Figure 1: Original image

Figure 2: Sobel operator applied

IV. POSE RECOGNITION

The proposed feature extraction and similarity measurement is divided into two steps: (1) feature extraction, (2) similiarity measurement using fuzzy local information c-means clustering algorithm.

A. GRAY SCALE CONVERSION

The obligatory images are rehabilitated into gray scale from RGB. as a result three bytes data (Red, Green and Blue components) is converted into single byte (Grayscale) data. This will aid in swift image processing such as edge detection and comparison.

B. COMPUTATION OF TRANSFORMATION MATRIX

A picture is denoted by **P**(*x,y*) where *x* can acquire values from *{h, l}* to signify a HighResolution or LowResolution picture correspondingly. The variable *y* can seize values from *{f, p}* depending on whether the pictures are in the frontal or a non frontal pose. The HR frontal picture are denoted by **P**(*h,f*)*i*, *i* = 1, 2, . . . , *N* and the LR non-frontal picture are denoted by **P**(*l,p*)*i*, where *N* is the number of picture. Correspondingly, **x**(*h,f*)*i* and **x**(*l,p*)*i* denote the analogous SIFT-based feature descriptors. The distance between the features **x**(*h,f*)*i* and **x**(*h,f*) *j* from the HR frontal picture is denoted by *d*(*h,f*)*i,j* . The goal is to simultaneously transform the feature vectors from $P_i^{(h,f)}$ and $P_j^{(l,p)}$ such that the Euclidean distance between the transformed feature vectors approximates the best possible distance $d_{i,j}^{(h,f)}$. To this end, we find the transformation W which minimizes the following objective function.

$$\mathbf{J}(\mathbf{W}) = \lambda \mathbf{J}_{\mathrm{DP}}(\mathbf{W}) + (1-\lambda)\mathbf{J}_{\mathrm{CS}}(\mathbf{W}) \quad (1)$$

The first term JDP is the distance preserving term which ensures that the distance between the transformed feature vectors approximates the distance $d_{i,j}^{(h,f)}$ and is given by

$$\mathbf{J}_{\mathrm{DP}}(\mathbf{W}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(q_{ij}(\mathbf{W}) - d_{i,j}^{(h,f)}\right)^2 \quad (2)$$

$$q_{ij}(\mathbf{W}) = \left|\mathbf{W}^T\left\{\phi\left(\mathbf{x}_i^{(h,f)}\right) - \phi\left(\mathbf{x}_j^{(l,p)}\right)\right\}\right|$$

is the distance between the transformed vectors of the picture $I_i^{(h,f)}$ and $I_j^{(l,p)}$. The second term of the objective function J_{CS} is an optional class separability term to further facilitate discriminability. We use a simple class preserving term that tries to minimize the distance between feature vectors belonging to same class and is of the form

$$\mathbf{J}_{\mathrm{CS}}(\mathbf{W}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\delta(\omega_i,\omega_j)q_{i,j}^2(\mathbf{W}) \quad (3)$$

where δ(ωi, ωj)=0 when ωi = ωj and 1 otherwise. Here ωi denotes the class label of the i^{th} picture. Clearly, the distance $q_{i,j}(W)$ and thus the objective function depends on the transformation matrix W. The relative effect of the two terms in the objective function is controlled by the parameter λ. Separating the terms containing W, the final objective function takes the form

$$\mathbf{J}(\mathbf{W}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{i,j}\left(q_{i,j}(\mathbf{W}) - \beta_{i,j}d_{i,j}^{(h,f)}\right)^2 \quad (4)$$

$$\alpha_{i,j} = (1-\lambda)\delta(\omega_i,\omega_j) + \lambda \text{ and } \beta_{i,j} = \lambda/\alpha_{i,j}.$$

Subsequently the iterative majorization algorithm is used to diminish the objective function to decipher the transformation matrix W. The fundamental idea of the majorization method is to swap iteratively the original function J(W) by an auxiliary function g(W;V). The auxiliary function, also known as the majorization function of J(W) is simpler to curtail than the original function. Please refer to [14] for minutiae of the algorithm.

C. CLUSTERING AND MATCHING SIMILIARITIES USING FLICM ALGORITHM

Clustering is a method pro classifying substance or patterns in such a method to facilitate samples of the identical cluster are more related to one another than samples belonging to different clusters. There are two main clustering strategies: the hard clustering system and the fuzzy clustering system. The conservative hard clustering methods classify each point of the data set just to one cluster. Though, in many real situations, issues such as restricted spatial resolution, deprived contrast, overlapping intensities occurs. Fuzzy set theory has introduced the system of biased membership, described by a membership task. It is a squashy segmentation method, has been broadly used, in image clustering and segmentation and it theater an significant task in solving evils in the areas of pose recognition and fuzzy model detection. The new-fangled feature should have some unique characteristics such as,

1) Integrate local spatial and gray level information in a fuzzy way to preserve robustness and noise insensitiveness.
2) Organize the persuade of the locality pixels depending on their distance from the central pixel.

3) Use the inventive image avoiding preprocessing steps that could cause detail missing. So, a novel fuzzy factor is introduced which is defined as,

$$G_{ki} = \sum_{\substack{j \in N_i \\ i \neq j}} \frac{1}{d_{ij}+1}(1-u_{kj})^m \|x_j - v_k\|^2 \tag{5}$$

Where the ***i*** th pixel is the centre of the local window, ***k*** is the reference cluster and the ***j*** th pixel belongs to the set of the neighbors falling into a window around the ***I*** th pixel (***Ni***). ***Dij*** is the spatial Euclidean distance flanked by pixels ***i*** and ***j***, ***ukj*** is the degree of membership of the ***j*** th pixel in the ***k*** th cluster, ***m*** is the weighting exponent on each fuzzy membership, and ***vk*** is the archetype of the center of cluster ***k***.

$$u_{ki} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|x_i - v_k\|^2 + G_{ki}}{\|x_i - v_j\|^2 + G_{ji}} \right)^{1/m-1}} \tag{6}$$

$$v_k = \frac{\sum_{i=1}^{N} u_{ki}^m x_i}{\sum_{i=1}^{N} u_{ki}^m}. \tag{7}$$

The FLICM algorithm is given as follows,
Step 1: Put the integer ***c*** of the cluster archetype, fuzzification parameter ***m*** and the stopping clause □ **.**
Step 2: Initialize arbitrarily the fuzzy divider matrix.
Step 3: Locate the loop counter ***b= 0.***
Step 4: Gauge the cluster prototypes using (1).
Step 5: Calculate membership values using (2).
Step 6**:** Next ***max {U (b) –U (b+1)}*** **<** □ then stop, or else, set ***b = b +1***and go off to step 4.
When the algorithm has converged, a defuzzification procedure takes consign in order to translate the fuzzy partition matrix ***U*** to a crusty partition. The utmost membership process is the most important method that has been developed to defuzzify the partition matrix ***U***. This procedure assigns the pixel ***i*** to the class ***C*** with the highest membership,

$$C_i = \arg_k \{\max\{u_{ki}\}\}, \quad k = 1, 2, \ldots, c. \tag{8}$$

The equation (8) is used to convert the fuzzy image achieved by the proposed algorithm, to the crusty segmented image.

$$J_m = \sum_{i=1}^{N} \sum_{k=1}^{c} \left[u_{ki}^m \|x_i - v_k\|^2 + G_{ki} \right]. \tag{9}$$

The measure used in the FLICM ideal function is still the Euclidean metric as in FCM, which is computationally easy. The difference between the original and FLICM applied image is described in figure 3 and 4.

Figure 3: Original image

Figure 4: Flicm algorithm applied

V. VIDEO AUTHORING AMID TEXT ANNOTATION AND SCENE STRUCTURED GRAPH

Scene Graph is high rank representation of a 3D world so as to be used to handle objects in a 3D graphics engine. Scene graphs canister, materialize in the formation of a text annotation. The chore of authoring multimedia is analogous to creating a text document by means of a word processing system. Depending on the features supported by the formatter, authors may be proficient to fluctuate the styling of the text, they may be able to diverge the spatial layout, and they may be able to integrate higher-level structures, such as chapters and sections. Similarly, multimedia authoring tools assign an author to assimilate numerous types of information into a composite presentation and depending on the system, It permit spatial and temporal

styling. Textual Annotation description tools provide different ways of creating textual annotation, from keyword- to linguistic-oriented annotation, based on who, what action, what object, where, when, why, and how. Unlike text, Image, video and audio assets can also used. Hence, different countries may use diverse written languages and thus using text may find obstacles in a multi-linguistic condition. Hence scene structured graph technique also applied in our manuscript, scene structured graph is a hierarchical structure containing nodes connected by edges. The nodes of the scene graph supervise the data describing a virtual scene and the edges that tie the nodes describe the relationships that subsist amid them in a evocative way. A scene graph organizes and controls the exposé of its ingredient objects. An imperative trait of all scene graph objects is that they can only be accessed or customized during the formation of a scene graph, except where overtly permitted. OpenGL Performer is an advanced API which is a low level API to render the scene and also it supports the Direct Acyclic model of the scene graph. The emerging standards of MPEG-7 provide a new process for authoring and presenting MPEG-7 Visual, standardizes the description tools we use to portray video and image content. The Visual descriptors are based on visual features that let us gauge similarity in images or videos. so, we can utilize the MPEG- 7 Visual Descriptors to seek out and sieve images and videos based on several visual features like color, texture, object shape, object motion, and camera motion. We can categorize the MPEG-7 Visual Descriptors into generic and high-level description tools. The generic Visual descriptors assent to illustrate color, texture, shape, and motion features. The high-level descriptors endow with description tools for face-recognition applications. The ultimate yield of the authoring process, we used MPEG-7. MPEG-7 Visual Descriptors into basic and sophisticated depiction tools. The generic Visual descriptors let us describe color, texture, shape, and motion features. The high-level descriptors afford depiction tools for face-recognition applications. So it gives efficient result based on the object we annotated. The following figure 6 is describe its result.

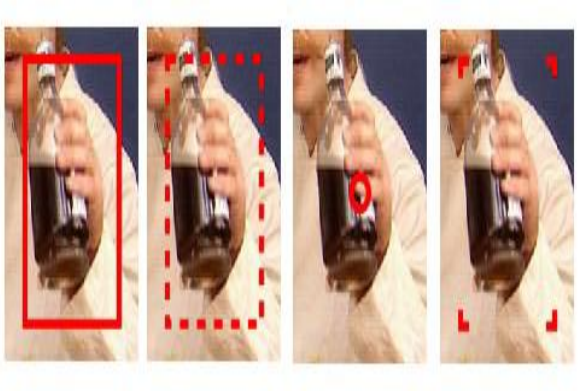

Figure 5. Authoring a video in MPEG-7

VI. CONCLUSION

Sketching is rampant at the design procedure. Hence in this paper, we proposed a sobel algorithm to sketch an image from the video for annotation and we also proposed a MDS-based approach for harmonizing LR facial images with substantial variations in pose and elucidation to HR images in frontal pose, then measures a similarity using Fuzzy Local Information C-Means clustering algorithm. In addition this paper comprises a dual method scene structured graph for video authoring.

References

[1]. Cui-Xia Ma, ”Sketch-Based Annotation and Visualization in Video Authoring” , IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 14, NO. 4, AUGUST 2012

[2]. Himanshu S. Bhatt, “On Matching Sketches with Digital Face Images” , IEEE TRANSACTIONS ON PATTERN ANALYSIS,

[3]. Khushboo Khurana,”Study of Various Video Annotation Techniques”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2013

[4]. Emily Moxley,”Video Annotation Through Search and Graph Reinforcement Mining”, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 12, NO. 3, APRIL 2010

[5]. I. Borg and P. Groenen, Modern Multidimensional Scaling: Theory and Applications. Springer, Second Edition, New York, NY, 2005.

[6]. B. S. Manjunath, P. Salembier, and T. Sikora, Eds.,” Introduction to MPEG-7: Multimedia Content Description Interface”. New York: Wiley, 2002.

[7]. S. L. Feng, R. Manmatha, and V. Lavrenko, “Multiple Bernoulli relevance models for image and video annotation,” in Proc. Computer Vision and Pattern Recognition (CVPR’04), 2004, pp. 1002–1009.

[8]. S. Prince, J. Warrell, J. Elder, and F. Felisberti, “Tied factor analysis for face recognition across large pose differences,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 6, pp. 970–984, June 2008.

[9]. S. Baker and T. Kanade, “Hallucinating faces,” in Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, March 2000.

[10]. ——, “Limits on super-resolution and how to break them,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 9, pp. 1167–1183, September 2002.

[11]. D. Bulterman and L. Hardman, “Structured multimedia authoring,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 1, no. 1, pp. 89–109, 2005.

[12]. F. Shipman, A. Girgensohn, and L. Wilcox, “Authoring, viewing, and generating hypervideo: An overview of hyper-hitchcock,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 5, no. 2, 2008, article no. 15.

[13]. B. W. Chen, J. C. Wang, and J. F. Wang, “A novel video summarization based on mining the story-structure and semantic relations among concept entities,” IEEE Trans.Multimedia, vol. 11, no. 2, pp. 295–312, 2009.

[14]. Y. J. Liu, K. L. Lai, G. Dai, and M. M. F. Yuen, “A semantic feature model in concurrent engineering,” IEEE Trans. Autom. Sci. Eng., vol. 7, no. 3, pp. 659–665, 2010.

[15]. Ivan Herman, “Graph Visualization and Navigation in Information Visualization: a Survey”, IEEE Trans.2000