Visual Sentences for Pose Retrieval Over Low-Resolution Cross-Media Dance Collections

Reede Ren, *Member, IEEE*, and John Collomosse, *Member, IEEE*

***Abstract*—We describe a system for matching human posture (pose) across a large cross-media archive of dance footage spanning nearly 100 years, comprising digitized photographs and videos of rehearsals and performances. This footage presents unique challenges due to its age, quality and diversity. We propose a forest-like pose representation combining visual structure (self-similarity) descriptors over multiple scales, without explicitly detecting limb positions which would be infeasible for our data. We explore two complementary multi-scale representations, applying passage retrieval and latent Dirichlet allocation (LDA) techniques inspired by the text retrieval domain, to the problem of pose matching. The result is a robust system capable of quickly searching large cross-media collections for similarity to a visually specified query pose. We evaluate over a cross-section of the UK National Research Centre for Dance's (UK-NRCD), and the Siobhan Davies Replay's (SDR) digital dance archives, using visual queries supplied by dance professionals. We demonstrate significant performance improvements over two base-lines: classical single and multi-scale bag of visual words (BoVW) and spatial pyramid kernel (SPK) matching [5].**

***Index Terms*—Content based image retrieval, dance archives, low-resolution pose similarity.**

I. Introduction

THE visual arts are increasingly turning to online digital archives for dissemination. Large online archives now exist for paintings and textiles (VADS), film (BFi ScreenOnline), and in recent years dance. Two examples are the Siobhan Davies Replay (SDR) and UK National Resource Centre for Dance (UK-NRCD[1]) have launched online dance archives. As additional dance collections appear in isolation online, a need emerges for a single portal enabling search across collections and across media types. The Digital Dance Archives (DDA) project recently launched such a portal[2] spanning $\sim$ 100 years of dance history. While an online presence aids dissemination of this wealth of material, methods for discovering this content remain limited. Searches are typically text-based, focusing on authorship or time-location meta-data rather than annotation of the choreography itself. Systems such as Labanotation [22] exist to describe dance pose. Yet such expert annotation is costly and not scalable to large legacy archives. This paper describes one forthcoming aspect of DDA visual search: searching based on the posture (pose) of performers. Rather than explicitly estimating skeletal position (or "soft" probability distributions of body parts), we encode pose implicitly within a novel representation we dub "Visual Sentences". The Visual Sentence representation and its application to pose search forms the core contribution of this paper.

A large dance archive, such as the UK-NRCD's, comprises many dance collections each containing video and photographic footage of a specific dance company. Dance researchers are often restricted by their familiarity with only certain collections or works within a collection. Our motivation is to produce a search tool capable of identifying new connections between dance collections and across media types. Dance collections are often *cross-media*, containing not only video of performance but also photographic records "*Contact Sheets*" of the rehearsals that capture experimental choreography that led up to the final production. Pose is the essential unit of dance. The ability to match static imagery in Contact Sheets with instances of poses in a performance presents a new way to research the development and history of dance.

Our work is unique in its focus on low-fidelity, monocular cross-media dance collections and in the breadth of dance material considered—comprising four partially digitized collections circa 1920–1970 from the UK-NRCD archive—*Extemporary Dance Theatre*, *Revived Greek Dance*, *Natural Movement*, and *Ludmila Mlada* collections—and in addition, material sampled from the Siobhan Davies Replay (SDR) collection. These sources represent the two major U.K. online dance archives at the time of writing. Although photographic content in these collections has been scanned at high resolution, the digitized video was originally captured on CineFilm or VHS tape and is grainy, and of low-resolution. Furthermore, the footage is typically shot with a monocular, fixed or panning camera from the back of a darkened theatre with frequently changing lighting conditions. This can result in contrast bleached, blurry edges and a typical performer height of only $\sim 50 - 100$ pixels.

Human pose estimation (HPE) is the process of explicitly estimating either a skeletal pose, or a probability map indicating the likelihood of skeleton limb positions. HPE has been demonstrated over a variety of broadcast and movie quality monocular footage [4], [10], [16], [27]. Fig. 1 illustrates the challenge of applying two HPE approaches claiming generality of input [4], [13] to the low fidelity dance footage considered in this paper. Although these approaches are robust to clutter, this is not an

Manuscript received September 10, 2011; revised January 13, 2012 and April 19, 2012; accepted April 28, 2012. Date of publication May 17, 2012; date of current version November 22, 2012. This work was supported under the AHRC "Digital Dance Archives" project AH/H037926/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhu Liu.

R. Ren is with the Department of Computer Science, University of Glasgow, Glasgow G12 8QQ, U.K. (e-mail: Reede.Ren@glasgow.ac.uk).

J. Collomosse is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Surrey GU2 7XH, U.K. (e-mail: J.Collomosse@surrey.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2199971

[1] National Resource Centre for Dance, University of Surrey, U.K.

[2] http://www.dance-archives.ac.uk

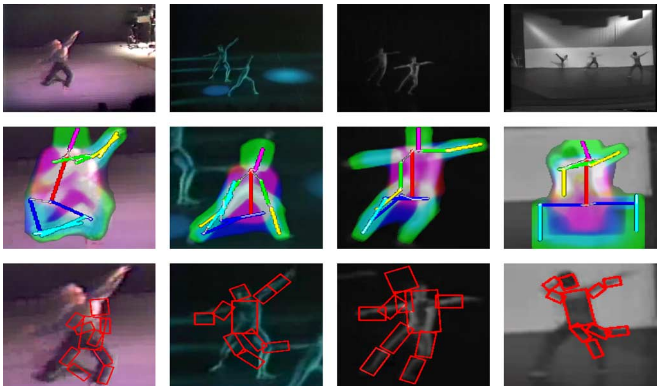

Fig. 1. Challenge of explicit HPE in our low-resolution archival dance footage. Performers are typically non-frontal facing, in diverse poses often $\sim 50-100$ pixels high (row 1). Even with their bounding box, localized HPE typically fails in these conditions; row 2 [11], row 3 [4].

issue in our footage; rather other conditions assumed by these methods (e.g., good contrast and resolution, limited self-occlusion) are violated. We therefore eschew explicit HPE, opting instead to implicitly encode pose through local spatial structure. Our pose representation is based on pixel information within a bounding box surrounding the performer. Although captured implicitly, we do not seek to label pixels as belonging to any particular part of the body.

Our core technical contribution is therefore a novel framework for pose retrieval on low fidelity dance footage, that *does not require explicit pose estimation*. We build upon Shechtman *et al.*'s self-similarity (SSIM) descriptor [33], computing descriptors at multiple scales within the bounding box. Descriptors are codebooked using hierarchical k-means and subsequently collected across scales to yield sequences of visual words—dubbed *visual sentences*—that collectively represent the content and scale-space containment of features within the bounding box (Section III).

Two complementary schemes inspired from the text retrieval literature—topic and passage analysis—are investigated to match the visual sentence representation, and compared against classical multi-scale approaches (Section IV). We evaluate our system over a corpus of dance footage in Section V using a query-set produced by dance professionals, omparing against search driven by explicit HPE [4], [13], and implict baselines for multi-scale matching, including classical BoVW and spatial pyramid kernel (SPK) matching.

II. Related Work

Most existing approaches to measuring pose similarity focus on human pose estimation (HPE) to extract 2-D or 3-D limb positions as a pre-process [14], [29]. HPE underpins many tracking and recognition applications, and has received broad attention for both single images [15] and video sequences captured from one [10] or multiple views [31].

HPE commonly begins with the localization of regions of interest containing people. The localization problem can be solved by background [38] or motion [2] subtraction, in simple cases. In cluttered scenarios, sliding window classifiers based on histogram of gradient (HoG) descriptors can robustly identify the torso [10], face [36], or entire body [8]. Following localization, pose estimation follows either: 1) top-down fitting of a person model, optimizing limb parameters and projecting to image space to evaluate correlation with image data; or 2) individually segmenting parts and integrating their positions in a bottom-up manner to produce a maximal likelihood pose.

Bottom-up HPE are becoming increasingly popular. Srinivasan and Shi [35] use graph-cut to parse a subset of salient shapes from an image and group these into a shape resembling a person. However the approach is limited to a single person, and clutter is reported to interfere with the initial segmentation. Mori and Malik sought to estimate 3-D pose by identifying the position of individual joints in a 2-D image. Scale and symmetry constraints were used to establish correspondence between a 2-D query image and training images annotated *a priori* with joint positions [27]. Ren *et al.* propose recursively splitting Canny edge contours into segments, classifying each segment as a putative body part using shape cues such as parallelism [30]. Ning *et al.* [16] apply a bag of visual words (BoVW) framework to learn codewords for body zone labelling—segmenting 2-D body parts to infer pose. We also apply BoVW in our work, but in the context of image retrieval [28], [34], where features are codebooked to form a representation used to form the basis of a search index.

Eichner and Ferrari presented "Pose Search" [13], a system using descriptors derived from the "soft" probability maps generated by their earlier HPE algorithm[10]. The system was refined for improved generality in [11], and operates over high (visual) quality broadcast TV footage: Buffy the Vampire Slayer. The approach relies on explicit HPE via pictorial structures [10]. A further approach to full-body HPE using pictorial structure was presented by Andriluka *et al.* [4], and is adaptable to pose search using a descriptor derived from estimated joint angles. In Fig. 1, and later in Section V-E, we show that these explicit HPE approaches to pose search are ill-suited to our diverse, low-resolution footage dance video shot at a distance.

Top-down approaches to 2-D HPE typically adopt a hierarchical limb model incorporating kinematic priors into their objective functions to bias toward feasible configurations. Lan and Huttenlocher [24] fit such a model using joint angles and considering the conditional independence of parts; more complex inter-limb dependencies (e.g., symmetry) are not considered. A more global treatment is proposed in [19] using linear relaxation, but shows good results only on uncluttered scenes. Most recently, pictorial structures [12] present a graphical model decomposing the objective function across edges and nodes in a tree. Spatio-temporal tracking of pictorial structures is applied to HPE in [23], and the fusion of pictorial structures with Ada-Boost based shape classification was more recent explored in [4]. Parallels between 2-D shape and 2-D pose retrieval are noted in [32] where local sensitive hashing (LSH) is used to retrieve pose in uncluttered scenes quite different to the low-fidelity archival footage we process.

Implicit representations for pose matching (i.e., omitting HPE) are relatively uncommon in the retrieval literature. Shechtman *et al.* proposed the SSIM descriptor [33]; essentially an autocorrelation surface computed local to a given key-point. SSIM demonstrates a degree of invariance across

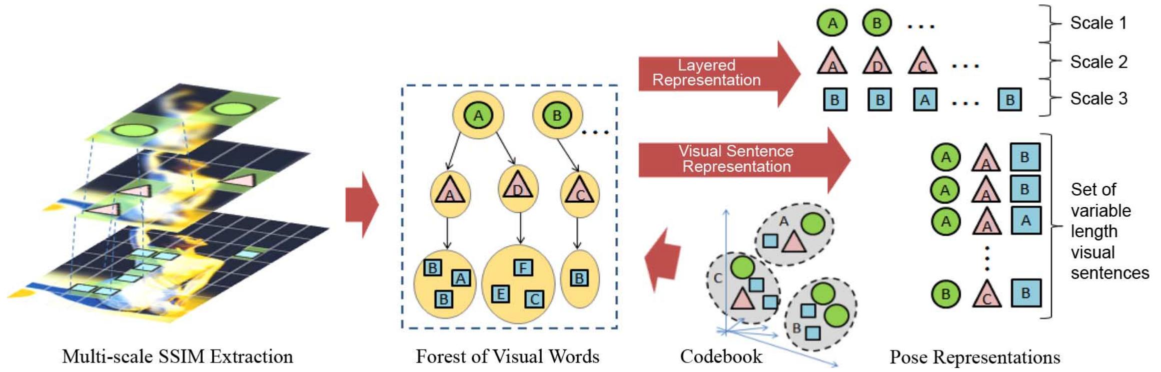
Layered Representation
Visual Sentence Representation
A
B
Scale 1
Scale 2
Scale 3
Set of variable length visual sentences
Multi-scale SSIM Extraction
Forest of Visual Words
Codebook
Pose Representations

Fig. 2. Computing the feature representation for pose. Self-similarity (SSIM) features [33] are computed over multiple scales. Scale space containment is captured through multiple tree representations, where nodes are features, and the root of each tree stems from features at the coarsest scale. Tree branches exhibit varying depth; not all regions yield a valid SSIM feature. The pose representation is formed by traversing the forest in a coarse to fine manner, collecting visual words derived form SSIM features across scales to yield ("visual sentences").

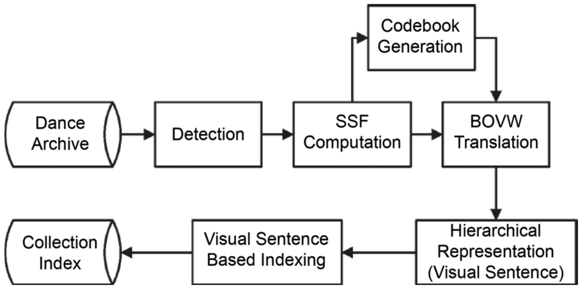

Fig. 3. Overview of the dance retrieval system; feature extraction (upper) is addressed in Section III; pose matching (lower) is addressed in Section IV.

depictive style, and can match shapes manifested in different textures. Most relevant to our work, pose retrieval is briefly demonstrated in [33] using an ice-skater clearly delineated against a white background. SSIM therefore formed a natural first step in our technique, but as we later show (Section V), straightforward application of [33] in a regular BoVW framework, does not deliver satisfactory performance over diverse archival dance footage.

III. Pose Representation

We now describe our system for searching dance footage using pose similarity, beginning with the (offline) process for extracting the pose representation. We then discuss the process for matching the pose representation in Section IV. Fig. 3 provides an overview of the entire system.

A. Performer Identification

We begin by identifying the region-of-interest (ROI) local to each dance performer or group of performers. Background subtraction based on motion [2] or color [38] is commonly used for ROI identification in biometrics or surveillance applications. Limited resolution inhibits the detection of repeatable interest points on the performers which frustrates motion flow based subtraction. Similar issues prevent camera motion compensation using features detected on the background. Consequently, we draw upon pedestrian detection algorithms, designed to operate over low-fidelity CCTV footage. We implement a variant of the Dalal and Triggs [8] pedestrian detector based on densely sampled HoG descriptors over four octave scale intervals. RBF-kernel SVMs are applied to identify neighborhoods likely to contain people. Neighborhoods are aggregated into rectangular ROIs using a heuristic optimization to minimize intersection while maximizing coverage [26]. A bounding box is obtained for each performer, or huddle of performers in the case of heavy inter-occlusion. These ROIs form our basic unit of retrieval.

B. Multi-Scale Feature Extraction

Our pose representation is based upon the SSIM descriptor [33], extracted at multiple scales over a low-pass pyramid within each bounding box. The bounded region is normalized by scaling to a regular 64×128 window; if the ROI is not of 1:2 aspect, then the shorter side of the box is padded to enforce this ratio. If the ROI is of greater than 1:1 aspect, then it is rotated 90° prior to conforming the aspect ratio and the subsequent feature extraction steps performed twice; once per 180° rotation of the ROI, to mitigate rotational ambiguity.

SSIM are computed across four scales at octave intervals using the luminance channel only. For computational convenience we up-scale the relevant level of the low-pass pyramid and compute SSIM densely at 5 pixel intervals, as follows. For a pixel $\mathbf{q}$, a small surrounding patch (5×5) is extracted and compared with overlapping patches forming a larger neighborhood (25×25) local to $\mathbf{q}$. The sum of squared differences (SSD) is calculated between patches. This results is a surface SSD_q local to $\mathbf{q}$ which is normalized and transformed into a correlation surface S_q:

$$S_q = exp\left(-\frac{SSD_q}{\max(\sigma^2_{noise}, \sigma^2_{auto}(q))}\right) \tag{1}$$

where σ^2_{noise} is a constant which reflects variance in SSD due to non-salient visual artifacts and $\sigma^2_{auto}(q)$ denotes the maximum "differential variance" within the 2×2 neighborhood local to q; these quantities are taken as originally defined by Shechtman *et al.* [33]. The surface S_q is then transformed into log-polar coordinates with origin at q. The space is quantized into 12 angular and 4 radial intervals. The maximum correlation values in each bin are collected to form a 48-dimensional SSIM descriptor local to q, which is then normalized. As in [33], a threshold on minimum and maximum values of $\sigma^2_{auto}(q)$ is employed to discard descriptors computed over poorly textured regions.

1) BoVW Codebook Generation: For each bounding box, the outcome of this process is a set of features in $\Re^{48}$ obtained through densely sampling descriptors over a regular grid at each scale of the low-pass pyramid. We quantize the space $\Re^{48}$ into k partitions via hierarchical k-means (HKM) clustering. The selection of an appropriate k is discussed in Section V-B. We adopt the classical or "hard assignment" BoVW strategy, as proposed by Sivic and Zisserman [34] mapping each feature to one of k codewords $\{\omega_1 \ldots \omega_k\}$ using nearest-neighbor assignment.

2) Stop-Word Removal: Stop-words are frequently occurring visual words corresponding to common components of pose. For example, a straight torso occurs frequently in dance. Codebooks containing visual words corresponding to this pose characteristic offer reduced discriminative power. Rather than prescribe such heuristics we identify stop-words through their frequency distribution; in information retrieval this is commonly achieved under a Bernoulli [3] or 2-Poisson [17] model. We adopt the former in our experiments (Section V-B2). The document frequency (DF) of each visual word is computed across the dataset, and expectation maximization (EM) used to estimate a Bernoulli distribution from the DF distribution. The probability of each visual word belonging to the distribution is estimated; visual words with probabilities below a constant "credit threshold" are deemed stop-words and discarded. Appropriate credit threshold values are explored in Section V-B2.

C. Hierarchical Representation

Defining the low-pass pyramid at octave scales enforces unique containment of a feature at one given scale within the footprint of a feature at a coarser scale. Thus a set of trees (forest) representation can be produced, with nodes corresponding to visual words (features), and connectivity defined by scale-space containment (Fig. 2, left). Each branch may be of variable length, since it is possible that an SSIM feature at a given scale is invalid, due to $\sigma^2_{auto}(q)$, inhibiting the discovery of descriptors at finer scales.

We now describe the pose representations considered in this work, derived from this multi-scale representation. The "visual sentence" representation underpins our novel contribution, whereas the "layered" representation serves as a baseline for performance comparison.

1) Visual Sentences: We propose the collection of visual words through traversal of each tree in the forest, collecting words along all root-leaf paths. This forms a set of visual word strings we term *visual sentences*. Each visual sentence encodes individual fine-scale regions, along with supporting context encoding local spatial relationships at coarser scales (Fig. 2). Visual words within the visual sentences are ordered left-right from coarse to fine scale. The sentences are right-padded with a stop-word token ω_{k+1} should they be of length less than four (corresponding to the four scales of the pyramid). The set of visual sentences forms our pose representation. Although relative spatial information is not explicitly coded in this (unordered) set, it is implicitly captured in the local contextual information within each visual sentence.

2) Layered Representation: Visual words can be encoded without context by simply collating into groups at each layer of the pyramid. We compute a normalized frequency histogram of word occurrence at each spatial scale. Such histograms may be treated independently during matching. For example, by assigning different weights to measures of similarity at different scales. We refer to the sequence of such histograms computed across each pyramid layer as the *layered representation* of the ROI. Such a representation is reminiscent of the multiple-scale collation of HOG descriptors computed by Pyramid-HOG (PHOG) [5]. Alternatively the scale-dependent histograms may be summed prior to normalization, and then globally normalized to yield a single frequency histogram. In this case the system degenerates to a standard hard-assignment BoVW approach, using dense-sampled SSIM as the descriptor.

IV. Pose Relevance Estimation

We now describe the process for matching a pair of pose representations, so establishing the relevance of an ROI or "document" d within the dataset $\mathcal{D}$, given a query ROI Q. We develop several "pose similarity functions" $(PSF1-5)$ using the layered $(PSF1-2)$ and visual sentence $(PSF3-4)$ representations, evaluating the performance of each in Section V.

Our retrieval process estimates $p(Q|d)$ for all $d \in \mathcal{D}$ and returns these documents as a ranked list of results in descending order of this probability. The user queries the database by selecting a performer. This is identified by drawing a bounding box on a single image or video frame. The drawn box is "snapped" to the spatially closest ROI detected during pre-processing, thus $Q \in \mathcal{D}$ obviating the need for any feature extraction at query-time.

We propose five pose similarity functions $(PSF1-5)$ potentially suitable for matching our hierarchical representation. $PSF1-2$ match the *layered representation* and serve later as a baseline for comparison to the state of the art Section V. $PSF3-4$ match the *visual sentence* representation, using two complementary approaches that propose handling of the spatial relationships between SSIM features in different ways. We also propose $PSF5$, a mechanism for enhancing diversity within the ranked results, to mitigate many similar video frames being returned by the system for a given pose. The relative performance of all similarity functions are compared in Section V.

A. Classical Multi-Scale BoVW (PSF1)

In the simplest case of the *layered* representation, features from all spatial scales can be collected and a single BoVW codebook formed over the entire feature set. $p(Q|d)$ is measured directly as the joint probability of all visual words in d occurring within Q:

$$p(Q|d) = \prod_{i=1}^{n} p(\omega_i|d). \tag{2}$$

where ω_i is the ith of n visual words present in Q. This is the classical BoVW "hard assignment" model as proposed by Sivic and Zisserman [34], trivial to compute as the pose representations are already frequency distributions of visual words $\omega_{1..k}$, i.e.,

$$p(\omega_i|d) = \frac{tf(\omega_i, d)}{|d|} \tag{3}$$

where tf indicates term frequency, and $|d|$ normalizes for document length (i.e., word count in the image).

Following the work of Nister and Stewenius [28], efficient feature lookup is implemented using an inverse index. In practice, for large diverse datasets such as UK-NRCD, a performance increase can also be obtained by blending $p(\omega_i|d)$ with the frequency of occurrence of ω_i within $\mathcal{D}$, i.e., $p(\omega_i|\mathcal{D})$ to improve robustness:

$$\hat{p}(\omega_i|d) = \lambda\frac{tf(\omega_i, d)}{|d|} + \frac{(1-\lambda)}{|\mathcal{D}|}\sum_{a\in\mathcal{D}} p(\omega_i|a). \tag{4}$$

In our comparative evaluation we set a blending factor of $\lambda = 0.8$. We refer to the pose similarity function utilizing metric $\hat{p}(\omega_i|d)$ as $PSF1$. $PSF1$ provides both a performance baseline for evaluation (i.e., classical BoVW [34]), and forms a component of the other pose similarity functions we now propose.

B. *Layered Representation (PSF2)*

Dance formalisms such as Laban's Choreutics [22] describe pose in terms of a hierarchy of body zones, and it is natural to consider pose similarity in these terms. However, as noted in Section I, for our dataset the explicit identification and analysis of body parts is infeasible. In the simple case of the layered representation, we consider our four scales to *directly* map to four granularities of body zone. Rather than combining our features into a mixed-scale BoVW as in $PSF1$, we individually weight the importance of each scale layer:

$$p_(Q|d) = \sum_{l=1}^{5} \eta_l \mathrm{PSF1}(Q|D;l) \tag{5}$$

where $PSF1(.)$ indicates similarity of visual words at layer l, under similarity function $PSF1$ (2). We train an SVM to learn appropriate weights $\eta_{1..5}$ using 50 manually constructed training queries and result sets sampled from $\mathcal{D}$. The resulting trained weights exhibit an approximately linear bias toward η at finer scales, identifying increased discrimination power at these scales. We refer to this weighted summation (5) score as $PSF2$.

C. *Topic Learning Over Visual Sentences (PSF3)*

Our *visual sentence* representation encodes both fine-scale features and their structural context; the spatial regions of the body, at increasing scales, within which each codeword occurs. In considering pose similarity, it is useful to consider such zones. However, since semantically meaningful body zones, e.g., an arm, or foot, are unlikely to map explicitly to the regions used in our scale-space hierarchy, it is better to consider the visual sentence as *implicitly* encoding body zone membership.

Body zones are therefore considered as latent variables or "topics". Each visual sentence is considered as having a membership distribution across the range of topics. This leads to a modification of $PSF1$:

$$p(Q|d) = \prod_{t=1}^{|\mathcal{T}|} p(t_i|d) \tag{6}$$

where individual topic presence with the unordered set of visual sentences (passage) is determined by

$$p(t_i|d) = \prod_{s\in S_d}^{|S_d|} p(s|t_i) \tag{7}$$

where S_d indicates the set of visual sentences within d, and $\mathcal{T}$ the set of topics. In practice, the $p(.)$ are computed using log-likelihoods and summed.

We draw upon latent dirichlet allocation (LDA) [9] to learn the underpinning topic distribution for our visual sentences. The identification of latent topics and their non-parametric distribution across document terms is a classical formulation of LDA, and we apply the Gibbs sampling approach of [18], learning the topic space by sampling 1 k documents from $\mathcal{D}$, and iteratively refining in alternation a set of topics $\mathcal{T}$ and their associated generative distributions for S_D. In our experiments we specify 48 topics to the LDA process. This is influenced by observations in Choreutics that indicate a similar number of atomic postures and zones of movement in dance [22]. Equation (6) reflects the similarity of topics present within the set of visual sentences, an approach frequently used in the analogous domain of text passage retrieval. We refer to this similarity measure as $PSF3$.

D. *Passage Retrieval Over Visual Sentences (PSF4)*

Spatial context over scale is implicitly encoded via the *visual sentence representation*; however $PSF3$ considers only the topic frequency distributions arising from these sentences rather than the spatial relationships between sentences. We propose a further similarity function $PSF4$ that explicitly encodes spatial groupings of visual sentences. We are again motivated by text retrieval literature that frequently compare local spatial context through so-called "passage analysis" to identify co-occurring phrases. The analysis of spatial co-located visual sentences is analogous in our system.

We define a 2 × 2 pixel sliding window w over the coarsest level of the low-pass pyramid extracted from the Q (i.e., 4×8 pixels). A similar sliding window is established over candidate d. The relevance $p(Q|d)$ is obtained by computing a similarity score $sim(.)$ between each pair of spatially corresponding windows in Q and d, respectively, averaged that score over the set of all valid window positions $\mathcal{W}$:

$$p(Q|d) = \frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}}^{n} sim(Q_w, d_w). \tag{8}$$

The similarity score between a pair of corresponding windows $sim(w, w')$ (typically containing $50 - 100$ visual sentences) is

Vocab.	PSF1		PSF2		PSF3		PSF4	
	(M/S BoVW)		(Layered/SVM)		(LDA/VS)		(Passage/VS)	
Size (k)	AP30	%	AP30	%	AP30	%	AP30	%
100	0.003	-	0.012	-	0.104	-	0.040	-
500	0.009	+200	0.030	+150	0.351	+238	0.181	+352
1000	0.013	+44	0.082	+173	**0.387**	+10	0.337	+86
2000	0.019	+46	0.190	+132	0.270	-30	0.325	-3
5000	0.032	+68	**0.220**	+16	0.206	-24	**0.347**	+7
10000	**0.034**	+6	0.185	-16	0.094	-54	0.249	-28

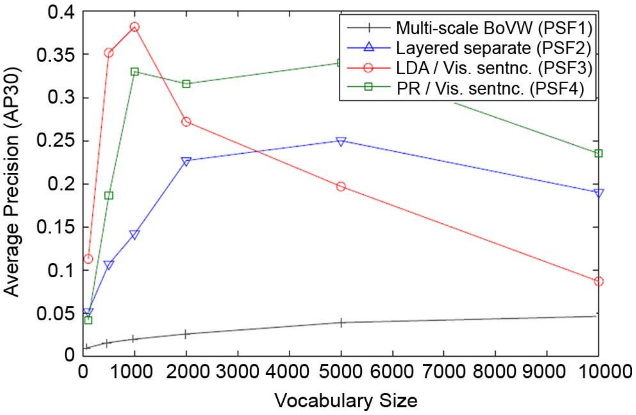

Fig. 4. Effect of vocabulary size (k) on performance for the layered representation at mixed ($PSF1$) and separate ($PSF2$) multiple scales, versus visual sentences using LDA with no re-ranking ($PSF3$) and Passage Retrieval ($PSF4$). In these experiments, no stop-word removal has been applied (see Tables I and II for analysis of best performing layered and visual sentence techniques with stop-word removal).

approximated by randomly sampling one third of visual sentences from each windows, and exhaustively searching for a pair of sentences in that subset minimizing:

$$sim(d, d') = \min_{i,j} \sum_{\min(|S_d|,|S_{d'}|)} \|s_i, s_j\| \qquad (9)$$

where $\|.\|$ indicates the count of in-place differences between visual words in sentence pair $\{s_i, s_j\}$.

E. Diversity and Re-Ranking (PSF5)

Direct presentation of results in strict order of similarity under (9) can lead to unsatisfactory grouping of results sampled from a narrow temporal window of one video. This lack of diversity (i.e., lack of dataset coverage at the expense of redundancy in results) runs counter to our aim of enabling cross-media and cross-collection discovery of dance.

Following recently proposed approaches in large-scale text retrieval we inject diversity into our results by re-ranking according to a clustering technique [7], [25] that discourages grouping of similar documents within the ranked results. Although re-ranking is possible over any of the four pose similarity functions defined here, we adopt $PSF3$, as this is later shown (Section V) to yield the most promising pose retrieval performance. We refer to $PSF3$ following this re-ranked process as similarity function $PSF5$.

We re-rank the top n results, according to the minimized intersection distance between one third of visual sentence pairs uniformly sampled from Q and d—using the measure of (9). Section V-C explores the choice of n (the scope of the re-ranking). A weighted affinity matrix $\mathbf{A}$ is computed between Q and all $\mathcal{D}$, based on (9). A matrix for each potential Q can be pre-computed offline, since $Q \in \mathcal{D}$. The resulting graph $\mathbf{A}$ is clustered using unsupervised Kruskal clustering [21], a greedy approach in which spanning trees are iteratively identified within the graph. The number of clusters is the number of extracted spanning trees, forming a cluster set $\{C_1, C_2, \ldots, C_n\}$, so that $\bigcap_{i \neq j}(C_i, C_j) = \emptyset$ and $\bigcup_{i=1}^{n} C_i = \mathcal{D}$. Documents within each cluster are then re-ranked independently under $PSF3$. The results presented to the user are based on a fusion of this independent ranked order, with tie-breakers resolved via the $PSF3$ score.

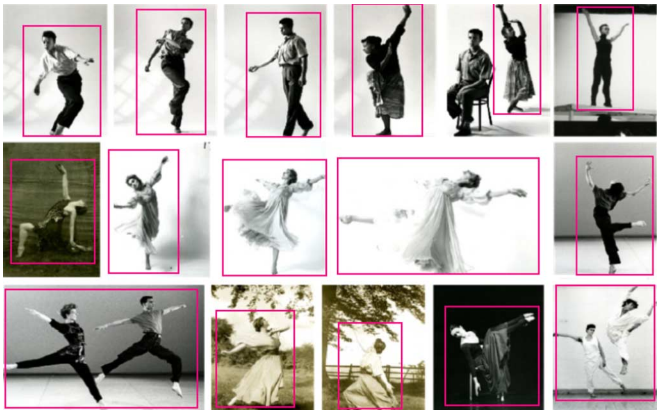

Fig. 5. Sample of the 65 pose search queries used in our evaluation.

V. Results and Discussion

We evaluated our system over two large-scale datasets. The first "core" dataset comprised 25 dance videos and 1684 Contact Sheet photographs. Sampling video key-frames every 5 s yielded $\sim 6.3k$ video stills, and so a total of around $6.5k$ images in the dataset. Approximately $8k$ bounding boxes (ROIs) were identified in the data. The source footage was sampled from 32 works across 4 cross-media collections of UK-NRCD footage: Extemporary Dance Theatre; Revived Greek Dance; Natural Movement; Ludmila Mlada. Videos were MPEG-2 PAL resolution footage (704x576) digitized from analogue CineFilm and VHS tape sources, and photographs were down-sampled to VGA resolution from high resolution archival scans of the original Contact Sheets with blurring applied to mitigate differences in fidelity between the photo and video media. This dataset is comparable in size to non-synthetic evaluation datasets of previous work focusing on *full-body* pose estimation, e.g., [16] and the Leeds Sport dataset [20], and an order of magnitude larger than [4] (but recall that we match pose only, and do explicitly estimate skeletal pose or soft pose maps). The second "extended"

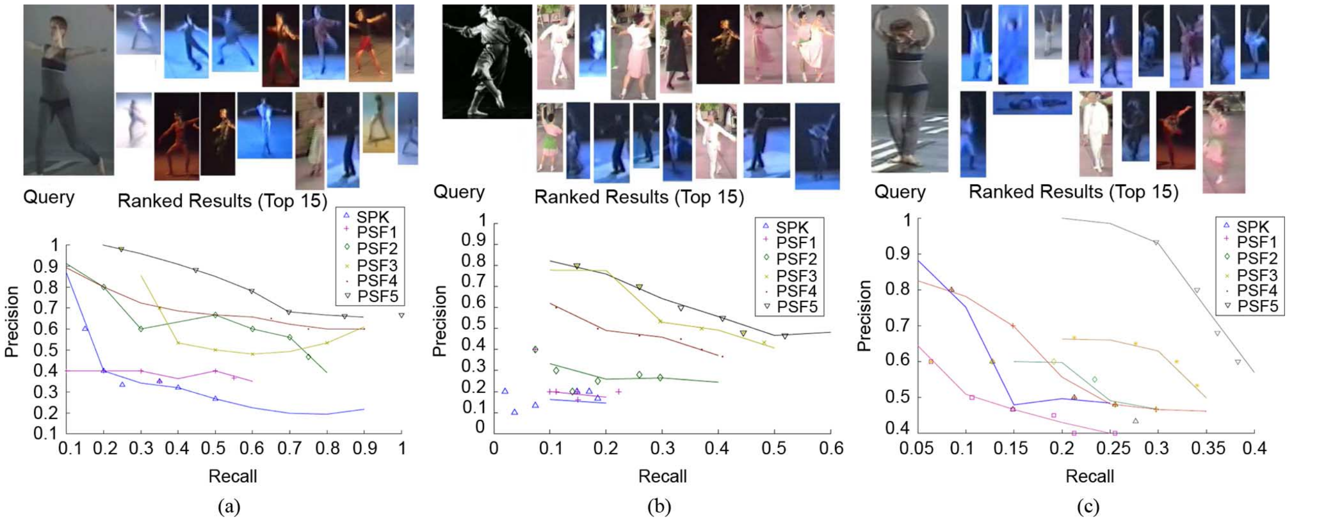

Fig. 6. Three representative retrieval results using the best performing system configuration $(PSF5)$. Top 15 ranked results are shown, ordered left-right, top-bottom. Associated Precision-Recall curves for all pose similarity functions $(PSF1-5)$ are plotted. Baselines for comparison are provided by classical multi-scale BoVW $(PSF1)$ and spatial pyramid kernel (SPK) matching.

dataset is a superset of the core dataset, including 200 additional videos and 542 production and Contact Sheet photographs from the Siobhan Davies Replay (SDR) collection. The SDR footage was similarly sampled yielding approximately $64k$ bounding boxes (ROIs), and so a total of $\sim 68k$ ROIs for this dataset. A similar order of scale-up also without a full annotation is used to test scalability in [4]. Due to the manual overhead in generating a ground truth, all experiments are reported using the core dataset, bar Section V-F.

An experimental GUI was developed to gather query ROIs from the core dataset. The queries comprised either key-frames from the videos, or monochrome contact sheets of rehearsals or experimental choreography leading to the production of that footage. We gathered 65 search queries from the UK-NRCD who identified poses in the video of choreographic interest (Fig. 5). Relevant images and video key-frames within the core dataset were also identified to create a ground truth. Fig. 6 provides sample queries with top 5 results.

A. Experimental Configuration

We measure performance using average precision computed over the top 30 ranked results; i.e., $P30$. Where a multiple query set $Q = \{q_1, \ldots, q_{65}\}$ is run in an experiment, this quantity is averaged over all 65 queries to give mean average precision (AP) at 30; i.e., $AP30$.

The five relevance estimation schemes proposed in Section IV are investigated, comparing the layered $(PSF1-2)$ and visual sentence $(PSF3-5)$ representations:

- $PSF1$: classic BoVW using SSIM over mixed scales.
- $PSF2$: BoVW using SSIM, run independently per scale, then fused.
- $PSF3$: latent dirichlet allocation (visual sentences).
- $PSF4$: passage retrieval (visual sentences).
- $PSF5$: LDA with re-ranking/diversity.

B. Visual Vocabulary

Visual vocabulary size is a key parameter in BoVW frameworks, governing the quantization level and thus expressivity of the codebook. Small vocabularies lack expressive power, while large vocabularies introduce excessive visual words resulting in poor discrimination.

1) Evaluating Codebook Size: We build BoVW codebooks by hierarchical k-means clustering of our extracted features. We compare the effects of our system using our layered $(PSF1, PSF2)$ and visual sentence $(PSF3, PSF4)$ representations without any re-ranking post-process or stop-word removal. Six vocabulary sizes are tried: $k = \{100,500,1000,2000,5000,10\,000\}$. Fig. 4 shows AP30 for $PSF2-4$ significantly outperforming classical BoVW $(PSF1)$ by at least $\sim 20\%$. Furthermore, visual sentences $(PSF3, PSF4)$ consistently outperform layered representations $(PSF1, PSF2)$ by around $\sim 10\%$. We conclude that: 1) independent multi-resolution analysis can generate more meaningful visual words than a single mixed resolution analysis; and 2) larger visual vocabularies between $1-5k$ yield the best retrieval performance. The results support multi-resolution approaches in a BoVW framework for dance pose retrieval.

Table IV (left) presents a general trend toward improved retrieval performance delivered with larger visual vocabularies up to around $k = 5000$—although performance improvement is not always linearly correlated with vocabulary size (k) especially in the case of the LDA based method $(PSF3)$. The LDA approach exhibits a 4% improvement in performance over the nearest rival $(PSF4)$, dropping sharply at $k = 1000$. This indicates LDA's inability to effectively learn topics using an over-complete BoVW codebook. The baseline of mixed resolution BoVW $(PSF1)$ is more sensitive to k than descriptors with separated resolutions, and suffers overall poorer performance. Visual sentences $(PSF3, PSF4)$ are more robust and reach the best performance with a smaller vocabulary size than the layered representation $(PSF2)$. This is likely due to the greater

TABLE I
EFFECT OF VARYING STOP-WORD THRESHOLD FOR THE BEST PERFORMING LAYERED SYSTEM, $PSF2$. B=BERNOULLI, P=POISSON

method(k)	0.7	0.75	0.8	0.85	0.9	0.95
B(1000)	0.193	0.214	0.22	**0.227**	0.223	0.212
P(1000)	0.192	0.23	0.235	**0.246**	0.24	0.22
B(5000)	0.223	0.235	0.246	0.25	**0.262**	0.254
P(5000)	0.222	0.232	0.24	0.248	**0.254**	0.254

TABLE II
EFFECT OF VARYING STOP-WORD REMOVAL THRESHOLD FOR BEST PERFORMING VISUAL SENTENCE SYSTEM (LDA NO RE-RANKING, $PSF3$). B=BERNOULLI, P=POISSON

method(k)	0.7	0.75	0.8	0.85	0.9	0.95
B(1000)	0.346	0.393	0.420	**0.422**	0.418	0.416
P(1000)	0.346	0.400	0.416	0.420	**0.422**	0.420
B(5000)	0.271	0.284	0.315	0.323	**0.327**	0.310
P(5000)	0.271	0.280	0.312	**0.322**	0.316	0.312

TABLE III
VISUAL SENTENCES. COMPARING THE PERFORMANCE OF LDA WITH THE RE-RANKING SCHEME ($PSF5$) VERSUS NO RE-RANKING ($PSF3$)

Re-ranking scope / PSF5	Precision(AP30)	% gain over PSF3
1	0.422	-
30	0.437	+3.6
60	0.453	+7.3
120	**0.487**	+15.4
150	**0.475**	+12.6

visual context captured across scales by the visual sentences. Performance increase slows after $k = 2000$ with degradation in performance at $k = 5000$. We therefore limit the scope of our later experiments to vocabularies of $1k$ and $5k$ visual words.

2) Stop Word Effectiveness: We cull stop words from the codebook (Section III-B2) by estimating the probability of a visual word belonging to the overall document frequency distribution, computed via either a Bernoulli or a 2-Poisson hypothesis. Tables I and II summarize the performance for the layered and visual sentence representations with different credit (removal) thresholds over these two vocabulary sizes. We observe no significant performance differences between adopting the Bernoulli or 2-Poisson hypotheses; however stop-word removal improves accuracy in $PSF3$ by around $\sim 6\%$. Bernoulli is adopted for convenience due to lower computational expense, and we observe AP30 to peak at a credit threshold of between 0.85–0.9, and we adopt 0.85 for stop-word removal in the remainder of our experiments.

C. Re-Ranking Performance

Based on the superior performance exhibited by $PSF3$ (Fig. 4) for $k \approx 1000$ we investigated re-ranking the top n ranked results (Section IV-E) to further improve accuracy ($PSF5$). Recall this process increases diversity (dataset coverage) by clustering results and executing a random walk local to each cluster. Table III illustrates the performance increase of $PSF5$ over $PSF3$ (i.e., no re-ranking; $n = 1$) under various re-ranking scopes n using a BoVW vocabulary of $k = 1000$. Re-ranking up to $n = 120$ offers an increase of around 6% (in relative terms, an increase of $\sim 15\%$) accuracy over the initial $PSF3$ results.

TABLE IV
LEFT: AVERAGE PERFORMANCE (AP30) OF EACH RETRIEVAL SCHEME ($PSF1 - 4$) COMPARED AGAINST BASELINES OF SINGLE-SCALE BOVW AND SPK MATCHING

Retrieval	k	
Scheme	1000	5000
PSF1 (Multi res. BoVW)	0.014	0.036
PSF2 (Layered/SVM)	0.223	0.254
PSF3 (LDA/VS)	**0.422**	0.273
PSF4 (Passage/VS)	0.343	0.357
PSF5 (LDA/VS/Re-Rank)	0.487	**0.288**
SPK	0.243	0.243
Single res. BoVW	0.019	0.032

D. Overall Performance and Baseline Comparison

We compare our proposed system against three baselines: 1) a classical hard-assignment BoVW framework using SSIM at a single scale [6]; 2) classical BoVW using a mixed bag of SSIM features computed over multiple scales (i.e., $PSF1$); 3) SPK matching [5]. These baselines are selected to evaluate the effectiveness of our hierarchical approach versus regular BoVW and a leading multiple resolution approach.

Table IV compares the performance (AP30) of all proposed schemes to the baselines, with precision-recall graphs for representative queries given in Fig. 6. SPK performance is comparable to that of $PSF1$ (multi-scale BoVW), but achieves around half the performance of our best performing similarity function ($PSF3$). Combining relevance across multiple scales ($PSF2$) yields fractionally improved results over SPK. All techniques significantly outperform single-scale BoVW using SSIM, which performs very poorly with apparently random results. These results support the adoption of multi-scale analysis for pose retrieval.

Both $PSF3$ and $PSF4$ outperform all baselines, indicating that the structuring of codewords across scale (i.e., through the visual sentence) is critical to improving precision. Although our passage retrieval approach ($PSF4$) performs most consistently as codebook size (k) varies, method $PSF3$ exhibits around 8% improvement in performance at $k = 1000$ outperforming all other approaches. The inclusion of the re-ranking operation on $PSF3$ (i.e., $PSF5$) produces a further improvement, resulting in a mean average precision of 48.7% for our final system (averaged over the 65 query set). We suggest that the improvement delivered via re-ranking may be due to the diversity of the dataset and the potential for ambiguity in the low-resolution pose query.

Three representative retrieval results are presented in Fig. 6, for the best performing retrieval scheme $PSF5$. Precision recall graphs are presented for all multiple resolution retrieval schemes (SPK and $PSF1 - 4$), showing visual sentence retrieval schemes ($PSF3-5$) to outperform the layered representation based schemes and SPK in all cases. In all cases $PSF5$ (i.e., $PSF3$ with re-ranking) exhibits best performance, with plain $PSF3$ a close second in two of three cases. Fig. 7 explores the failure cases where a pair of performers is mistakenly grouped into a single ROI by the performer detection stage. The system is able to retrieve another mistakenly grouped pair of performer ROIs in similar poses, and within the top 5 results returns poses broadly similar to those of the two performers.

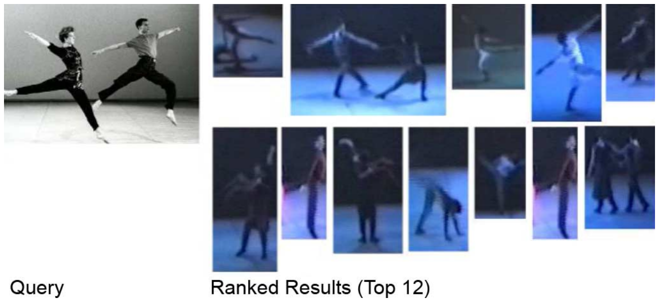

Fig. 7. Serendipity: multiple performers are accidently grouped within single ROI. However when posed as a query, another mistakenly grouped pair of performers is retrieved in a similar pose.

TABLE V
COMPARATIVE OF EXPLICIT POSE ESTIMATION

Precision	PSF3	PSF5	PSR	PS
AP30	0.323	**0.337**	0.054	0.201
AP60	0.338	**0.355**	0.002	0.163
AP90	0.270	**0.352**	<0.001	0.083

E. Implicit versus Explicit Pose Estimation

Prior approaches to full-body pose search explicitly estimate the location of body parts in the form of a probability density map [10], [11] which may be converted into a pose descriptor for search [13] or resolved to a single best estimate skeletal pose [4], [10], [11]. The joint angles of the skeleton could alternatively be used a descriptor. We now compare the proposed system against two techniques that explicitly estimate pose: Pictorial Structures Revisited (PSR) [4] and Pose Search (PS) [13]. For PSR we parse the output of this code to determine joint angles to yield a descriptor. This is necessary as PSR [4] stops short of proposing a pose search system *per se*. For PS we use the authors code from [13] to convert the soft pose estimates of [11] into a pose descriptor. We compare against PSF3 which exhibits the best performance without the diversity re-ranking step (PSF5), unavailable to PSR/PS (Table V).

Although these approaches function very well over the quality of footage for which they were designed, they perform pooly on lower resolution footage. To explicitly detect body parts these methods need higher contrast images with visible limbs of sufficient size. The higher performance of PS versus PSR is likely due to matching using soft pose estimates.

F. Extended Evaluation

Our experiments so far have focused on the large corpus of dance footage defined as our “core” database. We have run additional experiments on the “extended” dataset incorporating additional video material from the SDR collections. Ground truth markup is impractical given the size of this dataset; rather we measure result precision only, by manually inspecting the correctness of poses returned using our set of 65 query poses. The precisions of the top 30, 60, and 90 ranked results are tabulated below for each similarity function. Although performance is around 8%–10% lower on average (and very slightly improved for SPK, and multiscale BoVW), we observe the relative performance of the similarity functions to mirror our earlier experiments with $PSF5$ exhibiting overall best performance (Table VI).

TABLE VI
EVALUATION OVER THE EXTENDED UK-NRCD/SDR COMBINED DATASET

Precision	SPK	PSF1	PSF2	PSF3	PSF4	PSF5
AP30	0.051	0.067	0.280	0.323	0.293	**0.337**
AP60	0.115	0.148	0.310	0.338	0.320	**0.355**
AP90	0.107	0.127	0.308	0.278	0.270	**0.352**

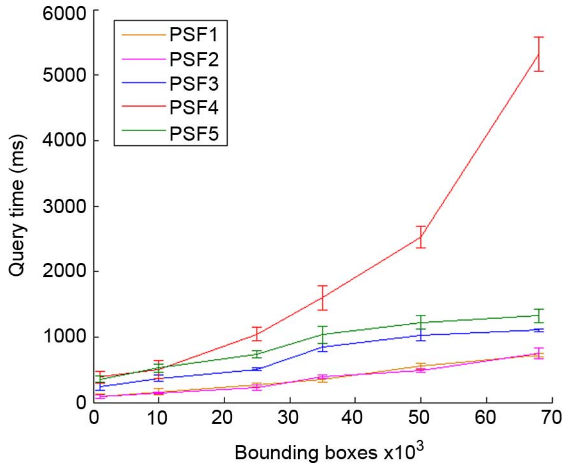

Fig. 8. Performance timings for PSF1–5 on $1k - 68k$ bounding boxes.

We have also performed timing analysis of approaches $PSF1 - 5$ computing mean query time (MQT) over the set of 65 queries. Fig. 8 indicates sub-linear scaling of $PSF1 - 3$ and $PSF5$ with respect to database size. In the case of $PSF5$ the diversity modelling adds a small approximately constant overhead to the processing times to $PSF3$. For our most successful approach $PSF5$, queries over the “core” dataset of $\sim 8k$ bounding boxes yielded an MQT of $1040ms \pm 65$, and over the “extended” dataset of $\sim 68k$ bounding boxes an MQT of $2640ms \pm 102$.

VI. CONCLUSION

We have presented a novel representation and matching scheme for pose retrieval over cross-media (photo and video) dance archives. Our system is designed to operate over low-resolution archival dance footage, exhibiting diverse content and artifacts arising from digitization noise, blur, and adverse illumination. We eschew explicit estimation of limb positions for implicit encoding of pose, delivered through the SSIM derived visual sentence representation. Our most successful matching scheme $(PSF3)$ harnesses LDA to learn topics within the dataset corresponding to body zones in canonical poses, and can be considered to build upon the hyperfeature meta-framework proposed by Agarwal and Triggs [1]. This approach to pose search using implicit pose representation has been shown to outrank classical BoVW using single and multiple scales, and a leading multi-resolution approach (spatial pyramid kernel). Further, our re-ranking post-process over $PSF3$ $(PSF5)$ delivers additional precision and introduces diversity into the resulting poses that promotes user discovery of content with the archive. Although explicit pose estimation is not our goal, we have presented illustrative comparisons of our implicit pose matching against two well-established baselines for full body pose estimation. A further project could explore extensive comparisons including recently proposed flexible part-based techniques [37].

A number of components in our system might be extended in future work. Further refinement is needed to robustify against view angle change and self-occlusion. Our use of a pedestrian detection algorithm [8], trained on dance footage, for performer localization averages only around $\sim 43\%$. This causes many ROIs to be absented from the index, prior to application of the pose matching techniques introduced in this paper. This lower than expected performance is likely due to wide variations in pose and appearance with the footage.

Acknowledgment

The authors would like to thank H. Roberts, S. Maxwell, and A. Reynolds (UK-NRCD); R. Fensham (U. Surrey); S. Whatley, D. Bennett (U. Coventry, SDR) for discussions, archive access, and annotation. The authors also would like to thank the CALVIN group for assistance with the comparative evaluation.

References

[1] A. Agarwal and B. Triggs, "Hyperfeatures—Multilevel local coding for visual recognition," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2006.

[2] A. Aggarwal, S. Biswas, S. Singh, S. Sural, and A. Majumdar, "Object tracking using background subtraction and motion estimation in MPEG videos," in *Proc. Asian Conf. Computer Vision (ACCV)*, 2006, vol. 3852.

[3] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Trans. Inf. Syst. (TOIS)*, vol. 20, no. 4, pp. 357–389, 2002.

[4] M. Andriluka, S. Roth, and B. Schiele, "Pictoral structures revisited: People detection and articulated pose estimation," in *Proc. Computer Vision and Pattern Recognition*, 2009.

[5] A. Bosch, A. Zisserman, and X. Muñoz, "Representing shape with a spatial pyramid kernel," in *Proc. CIVR*, 2007, pp. 401–408.

[6] K. Chatfield, J. Philbin, and A. Zisserman, "Efficient retrieval of deformable shape classes using local self-similarities," in *Proc. ICCV Workshops*, 2009, pp. 264–271.

[7] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *Proc. ACM SIGIR*, 2008.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. Computer Vision and Pattern Recognition*, vol. 3, pp. 886–893, 2005.

[9] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 91, pp. 993–1022, 1990.

[10] M. Eichner and V. Ferrari, "Better appearance models for pictorial structures," in *Proc. British Machine Vision Conf. (BMVC)*, 2009.

[11] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, Articulated Human Pose Estimation and Search in (Almost) Unconstrained Still Images, 2010, Tech. rep., ETH Zurich, D-ITET, BIWI, TR No. 272.

[12] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object detection," *Int. J. Comput. Vis.*, vol. 61, 2003.

[13] V. Ferrari, M. Martin-Jimenez, and A. Zisserman, "Pose search: Retrieving people using their pose," in *Proc. Computer Vision and Pattern Recognition*, 2009.

[14] D. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understand.*, vol. 1, no. 73, pp. 82–92, 1999.

[15] P. Guan, A. Weiss, A. Balan, and M. Black, "Estimating human shape and pose from a single image," in *Proc. Int. Conf. Computer Vision*, 2009.

[16] Y. G. H. Ning, W. Xu, and T. Huang, "Discriminative learning of visual words for 3D human pose estimation," in *Proc. CVPR*, 2008.

[17] S. Harter, "A probabilistic approach to automatic keyword indexing, part i on the distribution of speciality words in a technical literature," *J. ASIS*, vol. 26, pp. 197–216, 1975.

[18] H. Ishwaran and L. James, "Gibbs sampling methods for stick-breaking priors," *J. Amer. Soc. Inf. Sci.*, vol. 96, no. 453, pp. 161–173, 2001.

[19] H. Jiang, "Human pose estimation using consistent max-covering," in *Proc. Int. Conf. Computer Vision*, 2009.

[20] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. British Machine Vision Conf.*, 2010.

[21] J. Kleinberg and E. Tardos, *Algorithm Design*. Reading, MA: Addison-Wesley, 2005, pp. 158–161.

[22] R. Laban, *Choreutics*. London, U.K.: Macdonald and Evans, 1966.

[23] X. Lan and D. Huttenlocher, "A unified spatio-temporal articulated model for tracking," *Proc. Computer Vision and Pattern Recognition*, vol. 1, pp. 722–729, 2004.

[24] X. Lan and D. Huttenlocher, "Beyond trees: Common-factor model for 2D human pose recovery," in *Proc. Int. Conf. Computer Vision*, 2005, vol. 1, pp. 470–477.

[25] Y. Lv and C. Zhai, "Positional relevance model for pseudo-relevance feedback," in *Proc. ACM SIGIR*, 2010, pp. 579–586.

[26] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," *Proc. Computer Vision and Pattern Recognition*, 2008.

[27] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," *Proc. Computer Vision and Pattern Recognition*, pp. 326–333, 2004.

[28] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2006.

[29] R. Poppe, "Vision based human motion analysis: An overview," *Comput. Vis. Image Understand.*, vol. 1108, pp. 4–18, 2006.

[30] X. Ren, E. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *Proc. Int. Conf. Computer Vision*, 2005, vol. 1, pp. 824–831.

[31] S. Samangooei and M. Nixon, "Performing content-based retrieval of humans using gait biometrics," *Multimedia Tools Appl.*, vol. 49, no. 1, pp. 195–212, 2010.

[32] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, vol. 2.

[33] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. CVPR 2007*, Jun. 2007, pp. 1–8.

[34] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. Int. Conf. Computer Vision*, 2003.

[35] P. Srinivasan and J. Shi, "Bottom-up recognition and parsing of the human body," *Proc. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[36] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 2, no. 57, pp. 137–154, 2004.

[37] Y. Yang and D. Ramanan, "Articulated pose estimation using flexible mixtures of parts," *Proc. Computer Vision and Pattern Recognition*, 2011.

[38] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," *Proc. Computer Vision and Pattern Recognition*, vol. 2, pp. 459–466, 2003.

Reede Ren (S'06–M'09) received the Ph.D. degree in multimedia information retrieval from the information retrieval group in the University of Glasgow, Glasgow, U.K., in 2008.

He is a research fellow in the National E-Science Centre of Scotland. His research interest includes multimedia data modelling, knowledge extraction, video summarization, and multimodal interaction analysis.

Dr. Ren is a member of BCS.

John Collomosse (M'09) received the Ph.D. degree from the University of Bath, Bath, U.K., in 2004.

He is a lecturer (asstistant professor) within the Centre for Vision Speech and Signal Processing (CVSSP) at the University of Surrey, Surrey, U.K. Prior to joining CVSSP, he held a lectureship for 4 years in the Department of Computer Science at the University of Bath. His research explores the enhancement of visual media collections through visual search (e.g., sketch, pose, or photo search) and artistic stylization.