Recognition in Ultrasound Videos: Where am I?

Roland Kwitt[1], Nuno Vasconcelos[2], Sharif Razzaque[3], and Stephen Aylward[1]

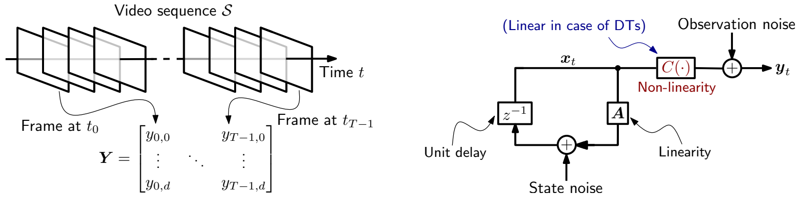
[1] Kitware Inc., Carrboro, NC, USA
[2] Dept. of Electrical and Computer Engineering, UC San Diego, USA
[3] Dept. of Computer Science, UNC, Chapel Hill, USA

Abstract. A novel approach to the problem of locating and recognizing anatomical structures of interest in ultrasound (US) video is proposed. While addressing this challenge may be beneficial to US examinations in general, it is particularly useful in situations where portable US probes are used by less experienced personnel. The proposed solution is based on the hypothesis that, rather than their appearance in a single image, anatomical structures are most distinctively characterized by the variation of their appearance as the transducer moves. By drawing on recent advances in the non-linear modeling of video appearance and motion, using an extension of dynamic textures, successful location and recognition is demonstrated on two phantoms. We further analyze computational demands and preliminarily explore insensitivity to anatomic variations.

1 Motivation

Many developing countries, as well as rural areas of developed nations, do not have immediate access to expensive medical imaging equipment such as MR or CT. As highlighted in a recent study [8], ultrasound (US) imaging is particularly well-suited for those underserved areas, since it is low-cost, versatile and non-invasive. Additionally, medical care in emergency vehicles and military field operations may benefit from US when performing first care. However, as emphasized in [8], training is needed for personnel to unfold the full potential of US imaging; yet this training requirement can be problematic in rural and emergency situations due to cost and circumstance. In addition to US interpretation, high US acquisition quality is essential but often difficult to achieve.

A cost-effective and practical solution to the training challenge is to embed expertise into the US system and/or a connected mobile device. This has led other groups to attempt to embed computer-aided diagnosis (CAD) systems into imaging devices. We instead seek to help an inexperienced operator acquire *clinically significant* images that can then be transferred to a central location for reading by experts. This approach should be easier to achieve and more broadly useful than an embedded CAD system. The technical challenge reduces to developing algorithms for recognizing key anatomic structures as the US videos are acquired. Based on those localizations, the system can then convey to the operator how to acquire additional images, relative to those key locations, for transmission to expert readers, or can indicate when US video must be reacquired to meet quality requirements. Apart from being helpful to novice users,



(a) Data matrix

(b) Generative model of the KDT

Fig. 1. (a) Assembly of the data matrix $\boldsymbol{Y}$ from a video sequence $\mathcal{S}$; (b) Generative model for the KDT.

we argue that automated recognition of anatomical structures might also be beneficial to experienced physicians, since it could help to minimize the risk of misidentifications.

The objective of this work is to promote a new pathway for locating anatomical structures when moving an US transducer. The key idea is to *avoid* image-to-image comparisons using an atlas but rather to exploit the full spatio-temporal information of the US video sequences. It is argued that the appearance changes of anatomical structures, due to probe motion, are particularly distinctive for their localization. Technically, we draw on recent advances in video modeling in computer vision. The varying appearance of an anatomical structure is represented by a generative video model, known as the *kernel dynamic texture* [2]. Similarity between video sequences is then defined as similarity in the parameter space of this model. Since we propose storing a database of key-location US sequences on portable devices and performing real-time analysis of US videos as they are acquired, generative models are particularly useful. In our case, we only need to augment the database by the KDT model parameters (which have a small memory footprint) and distances can be very efficiently computed.

While classification of US images has been previously studied (e.g., [7]), to the best of our knowledge, this is the first work to tackle localization on the basis of dynamic US sequence information. This paper presents 1) our application of the kernel dynamic texture algorithm, 2) a preliminary study on sensitivity and specificity using phantoms (admittedly for a limited range of the relevant problem space) and 3) a study on robustness towards simulated anatomic variations between the modeled structures to be localized and the actual observations.

2 Recognition with Kernel Dynamic Textures

We selected *dynamic texture (DT)* [5] models as an appropriate class of generative models for capturing video appearance changes. DT models arose from computer vision and were selected for US modeling because of the prominent role texture plays in US images, e.g., compared to edges or intensity. In particular, we exploited a recent non-linear extension of the DT family, denoted the *kernel dynamic texture (KDT)* [2], to capture non-linear appearance changes that will occur as structures move into and out of the ultrasound imaging plane.

Consider a US sequence $\mathcal{S}$ as an ordered sequence of T video frames, i.e., $\mathcal{S} = (\boldsymbol{y}_0, \dots, \boldsymbol{y}_{T-1})$, where $\boldsymbol{y}_t \in \mathbb{R}^d$ is the frame observed at time t. Under the DT framework of [5], these observations are modeled as samples of a *linear dynamical system (LDS)*. At time t, a vector of state coefficients $\boldsymbol{x}_t \in \mathbb{R}^T$ is first sampled from a first-order Gauss-Markov process, and the state coefficients are then linearly combined into the observed video frame $\boldsymbol{y}_t$, according to

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \tag{1}$$

$$\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{v}_t \tag{2}$$

where $\boldsymbol{A} \in \mathbb{R}^{T \times T}$ is the *state-transition* matrix and $\boldsymbol{C} \in \mathbb{R}^{d \times T}$ is the *generative* matrix that governs how the state determines the observation. Further, $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ denote state and observation noise with $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$, respectively. Assuming that the observations are centered[4] and following the system identification strategy of [5], $\boldsymbol{C}$ is estimated by computing an SVD decomposition of the data matrix $\boldsymbol{Y} = [\boldsymbol{y}_0 \cdots \boldsymbol{y}_{T-1}]$ as $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and setting $\boldsymbol{C} = \boldsymbol{U}$. The state matrix $\boldsymbol{X} = [\boldsymbol{x}_0 \cdots \boldsymbol{x}_{T-1}]$ is estimated as $\boldsymbol{X} = \boldsymbol{\Sigma}\boldsymbol{V}^\top$ and $\boldsymbol{A}$ can be computed using least-squares as $\boldsymbol{A} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_{T-1}][\boldsymbol{x}_0 \cdots \boldsymbol{x}_{T-2}]^\dagger$, where † denotes the pseudoinverse. When restricting the DT model to N states, $\boldsymbol{C}$ is restricted to the N eigenvectors corresponding to the N largest eigenvalues. The rest follows accordingly. Due to space limitations, refer to [5] for details on noise parameter estimation. In the non-linear DT extension of [2], the generative matrix $\boldsymbol{C}$ is replaced by a non-linear observation function $C : \mathbb{R}^T \to \mathbb{R}^d$, i.e.,

$$\boldsymbol{y}_t = C(\boldsymbol{x}_t) + \boldsymbol{v}_t, \tag{3}$$

while keeping the state evolvement linear. The corresponding dynamical system is denoted a *kernel dynamic texture (KDT)*, shown in Fig. 1(b). The non-linearity of C requires a different, although conceptually equivalent, set of parameter estimates. The idea is to use kernel PCA (KPCA) to learn the inverse mapping $D : \mathbb{R}^d \to \mathbb{R}^T$ from observation to state space, in which case the KPCA coefficients then represent the state variables.[5] We note that the KDTs are not necessarily restricted to work with intensity observation matrices; they will work with any kind of feature for which we can define a suitable kernel, c.f. [3].

Additionally, we have chosen to adopt the distance measure from [2] for measuring similarity of two video sequences, $\mathcal{S}^a$ and $\mathcal{S}^b$. This approach was chosen for its speed. It is based on an adaption of the Martin distance [6] among the corresponding DTs $\mathcal{D}_a = (\boldsymbol{A}_a, \boldsymbol{C}_a)$ and $\mathcal{D}_b = (\boldsymbol{A}_b, \boldsymbol{C}_b)$ with N states each. The (squared) Martin distance, given by [6,4]

$$d^2(\mathcal{S}^a, \mathcal{S}^b) = -\log \prod_{i=1}^{N} \cos^2(\phi_i), \tag{4}$$

is based on the subspace angles ϕ_i among the infinite observability matrices $\boldsymbol{O}_a$ and $\boldsymbol{O}_b$, defined as [4] $[\boldsymbol{C}_a^\top \; (\boldsymbol{C}_a\boldsymbol{A}_a)^\top \; (\boldsymbol{C}_a\boldsymbol{A}_a^2)^\top \cdots]^\top =: \boldsymbol{O}_a$. In fact, the $\cos(\phi_i)$

[4] Centering is straightforward by subtracting the column-wise means of $\boldsymbol{Y}$.

[5] See supp. material to [2] for centering in the feature space induced by the kernel.

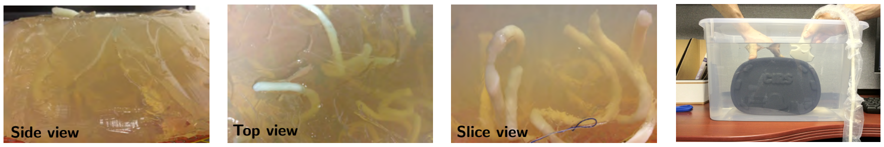


Fig. 2. Illustration of the *noodle phantom*, made of gelatine and Soba noodles (left three images) and an abdominal CIRS phantom mounted in a water tank (right).

correspond to the N largest eigenvalues λ_i of the generalized eigenvalue problem

$$\begin{bmatrix}\mathbf{0} & \boldsymbol{O}_{ab}\\ \boldsymbol{O}_{ba} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} = \lambda \begin{bmatrix}\boldsymbol{O}_{aa} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{O}_{bb}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \quad \text{with } \boldsymbol{O}_{ab} = \boldsymbol{O}_a^\top \boldsymbol{O}_b, \tag{5}$$

subject to $\boldsymbol{x}^\top \boldsymbol{O}_{aa}\boldsymbol{x} = 1$ and $\boldsymbol{y}^\top \boldsymbol{O}_{bb}\boldsymbol{y} = 1$. For DTs, computation of $\boldsymbol{O}_{ab}$ is straightforward, since the terms $\boldsymbol{C}_a^\top \boldsymbol{C}_b$ can be evaluated. For KDTs, it can be shown that computation of $\boldsymbol{C}_a^\top \boldsymbol{C}_b$ (which are no longer available) boils down to computing the inner products between the principle components of kernel matrix $K^a_{ij} = k(\boldsymbol{y}^a_i, \boldsymbol{y}^a_j)$ and $K^b_{ij} = k(\boldsymbol{y}^b_i, \boldsymbol{y}^b_j)$, i.e.,

$$\boldsymbol{O}_{ab} = \sum_{n=0}^{\infty} (\boldsymbol{A}_a^n)^\top \underbrace{\boldsymbol{C}_a^\top \boldsymbol{C}_b}_{\text{DTs}} \boldsymbol{A}_b^n \rightarrow \sum_{n=0}^{\infty} (\boldsymbol{A}_a^n)^\top \underbrace{\tilde{\boldsymbol{\alpha}}^\top \boldsymbol{G} \tilde{\boldsymbol{\beta}}}_{\text{KDTs}} \boldsymbol{A}_b^n, \tag{6}$$

where $\tilde{\boldsymbol{\alpha}} = [\tilde{\boldsymbol{\alpha}}_0 \cdots \tilde{\boldsymbol{\alpha}}_{T-1}]$, $\tilde{\boldsymbol{\beta}}$ are the (normalized) KPCA weight matrices with $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i - {}^1\!/\!{}_N(\boldsymbol{e}^\top \boldsymbol{\alpha}_i)\boldsymbol{e}$ and $\boldsymbol{G}$ is the kernel matrix with entries $G_{ij} = k(\boldsymbol{y}^a_i, \boldsymbol{y}^b_j)$. In the remainder of the paper, we use (4)-(6) for measuring similarity between US sequences and a standard RBF kernel for all kernel computations.[6]

For localization, we follow a sliding-window strategy, measuring how well a *key sequence* matches a subsequence from a long path (i.e., the *search sequence* $\mathcal{P}_n$) of acquisitions. That is, given Q frames in a key sequence, we move a sliding-window $\mathcal{W}_i$ of Q frames along a path by p frame increments. For each $\mathcal{W}_i$, we estimate the KDT parameters and compute the Martin distance to the KDT of the key sequence. A key sequence is indicated in a search sequence when the distance is minimal. At this time these minimums are illustrative. As more data and specific applications evolve, statistical likelihood methods will be used.

3 Experimental Protocol

For the studies in this paper, we use two different kinds of phantoms: 1) a homemade *noodle phantom* made of gelatine with embedded Soba noodles and 2) a triple modality 3D abdominal phantom (CIRS Model 057) mounted in a water tank, see Fig. 2. The noodle phantom is particularly useful, since the

[6] For KPCA, kernel width is set as $\sigma^2 = \text{median}_{i,j}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$; to compute G_{ij}, it can be shown [2] that $\boldsymbol{y}^a_i$ and $\boldsymbol{y}^b_j$ need to be scaled by σ_a and σ_b first.

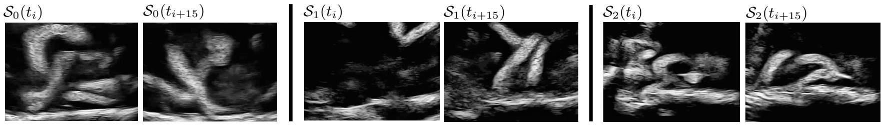

$\mathcal{S}_0(t_i)$
$\mathcal{S}_0(t_{i+15})$
$\mathcal{S}_1(t_i)$
$\mathcal{S}_1(t_{i+15})$
$\mathcal{S}_2(t_i)$
$\mathcal{S}_2(t_{i+15})$

Fig. 3. Snapshots of three key structures at two time points on the noodle phantom.

noodles are self-similar at a small scale, have ambiguous patterns of bends at medium scales, and at large scales and in US sequences present a rich set of structures that are difficult to casually distinguish.

For imaging we use the Telemed LogicScan 128 INT-1Z kit. US frequency is set to 5Mhz. Penetration depth is 90mm on the noodle phantom and 150mm on the abdominal phantom. *Speckle reduction* is enabled in the US acquisition software. All images were acquired freehand, without tracking. We learn $N = 5$ state KDTs and clip each sequence to a central 300×300 (noodle phantom), or 200×200 (abdominal phantom) pixel window. Using more states did not lead to any improvements in the presented results.

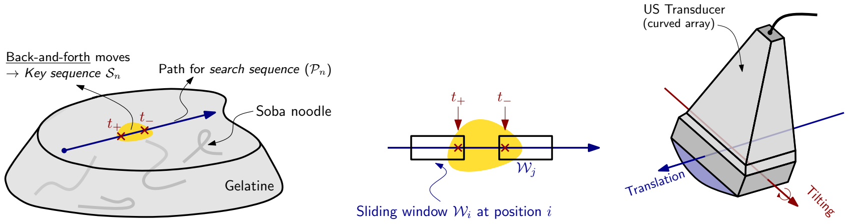
3.1 Localization of structures within US sequences

The first experiment tested whether it is possible to localize key structures in the noodle phantom. Two different sets of acquisitions were made. The first set was composed of short (40 frames) US *key sequences* $\mathcal{S}_n$, captured by moving the US transducer over three different key structures to be localized, see Fig. 3. Key structures were chosen ad hoc by the probe operator. We then estimated KDT models for each of the three key structures using this first set of data. The second set was composed of longer US *search sequences* $\mathcal{P}_n$, acquired along multiple paths on the noodle phantom; these simulated searches for the key structures, see Fig. 4. On both sets, we tried to minimize probe tilt and rotation, but rotation and titling was inevitable. Note also that the acquisition direction of the key sequences matched the acquisition direction of the search sequences.

To evaluate *sensitivity*, we performed the localization using key sequences applied to search sequences that also covered the corresponding key structures. Distance plots are shown on the left-hand side of Fig. 5. To evaluate *specificity*, we repeated this experiment along multiple search paths that did *not* cover any of the key structures. Distance plots are on the right-hand side of Fig. 5.

To evaluate the robustness against shifts of the ultrasound imaging plane (e.g., partial inclusion of a key structure), we performed ten runs with random displacements δ_x, δ_y of the clipping window in x and y direction with $\delta_x, \delta_y \in \{-5, \ldots, 5\}$ pixel. Fig. 5 shows the Martin distance averaged over all clipping window positions for each sliding window index along three search paths (left). The enclosing light-blue hull illustrates the standard deviation.

Based on the above three experiments we make the following observations: 1) key structures exist at global minima in the Martin distance metric of a search sequence, when key structures are encountered; 2) Martin distance decreases as



(a) Acquisition (b) Overlap (c) Movement

Fig. 4. Illustration of (a) the acquisition process on the noodle phantom, (b) sliding windows overlapping key structures (yellow) and (c) probe movements. In (b), hand-annotations t_+ and t_- bracket where the sliding window overlaps the key structure.

the sliding window moves towards a key structure and increases as it leaves the key structure; 3) if a key structure is not encountered by a search, then there is not a distinctive minimum in the distance measurements.

3.2 Localizing a hepatic vessel on an abdominal phantom

Our second experiment is more challenging in the sense that we try to locate a more subtle structure, namely a specific section of a *hepatic vessel* in an abdominal phantom. The experimental protocol is similar to the previous experiment; however, US transducer *tilting* (also known as angulation [1]) (see Fig. 4) is used instead of translation along a path. We attempt to localize the hepatic vessel key structure within a single search sequence. The key sequence acquisition spans $\approx 40°$ around the angle where the vessel is visible. The search sequence covers $\approx 140°$ around the vessel. Again, all acquisitions were performed freehand, and the ultrasound probe was repositioned on the phantom between each acquisition. Fig. 6 shows the Martin distances for localization and for localization using shifted clipping windows.

This experiment highlights two things. First, we can again localize the key sequence within the longer search sequence, even though the span of the minimal Martin distances that correspond to the *true* location of the vessel is less prominent and less persistent than in the previous experiment. Second, variation in the distance measurements is much higher for small vessels than for the more distinct, larger structures form the noodle phantom.

3.3 Localization in the presence of simulated anatomical variation

Our third experiment focused on the insensitivity of the distance metric and the localization method to anatomic variations. Specifically, we simulate anatomic variation by (non-linear) spatial distortion of the search sequences. We admit that this does not cover the range of variation among individuals, but it does begin to give an impression of robustness. Fig. 7 shows the distance measurements when the key structure is encountered within the search sequence. The

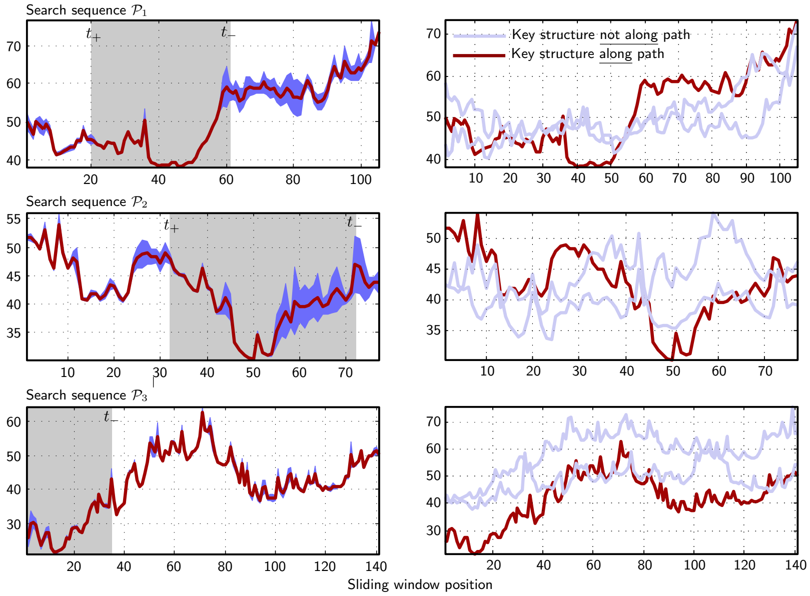


Fig. 5. Martin distance between key sequence KDT and the sliding window KTDs for three different paths, averaged over ten random clipping window positions (left); sliding window positions where the key structure is covered to some extent are marked light gray (from manual annotation); Distance measurements when trying to locate a key structure that was not covered by a path (right).

distortions are illustrated on a checkerboard pattern. Note that the *underwater* distortion is changing over time.[7] As shown, the distortions do not have a negative impact on the localization, although the *underwater* distortion leads to a less characteristic minimum.

3.4 Discussion and future work

We conclude that the KDT framework + Martin distance is an effective combination in localizing short US sequences of key structures. Further, computational requirements are modest, even for our un-optimized (MATLAB) code and use of intensity features. For example, given a key sequence consisting of 40 frames, computing the KDT model and distance metric per sliding window requires ≈ 0.1 seconds on an Intel Core i7 1.7Ghz CPU with 8GB of RAM. It is also worth noting that we can perfectly recognize the short US key sequences using a simple nearest-neighbor classifier and the Martin distance as a metric. For future work we note that using intensity information as our observations space has its limitations. Due to the generic nature of the KDT approach and KPCA-based system identification, we could easily integrate more specifically tailored US features as long as we can define a suitable kernel.

[7] See supp. video at http://vimeo.com/rkwitt.

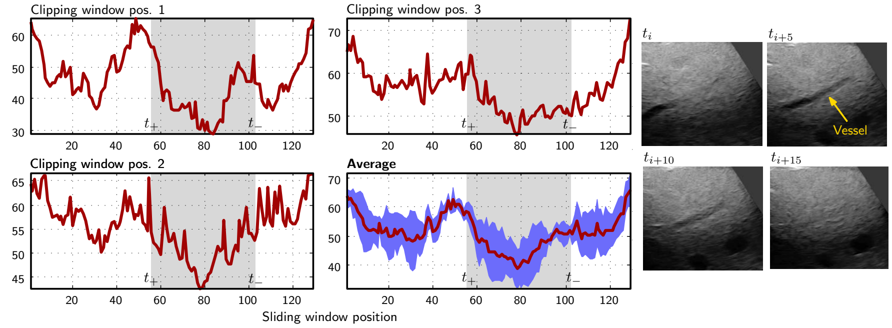


Fig. 6. Martin distance measurements for three clipping window positions and distance measurements averaged over all random runs (left); illustration of the hepatic vessel appearing and disappearing in the key sequence (right).

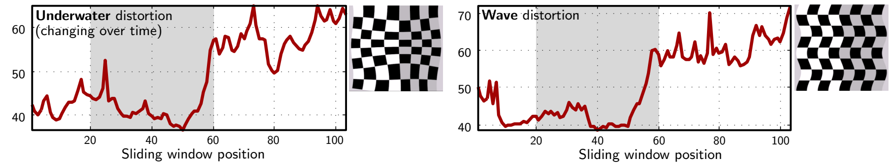


Fig. 7. Distortion experiments on the search sequence for two different types of (non-linear) spatial distortion (illustrated on the checkerboard pattern).

Acknowledgements This work was partially funded by NSF (CCF-0830535) and NIH/NCI (1R01CA138419-0, 2R44CA143234-02A1).

References

1. Block, B.: The Practice of Ultrasound: A Step-by-Step Guide to Abdominal Scanning. Thieme, first edn. (2004)
2. Chan, A.B., Vasconcelos, N.: Classifying video with kernel dynamic textures. In: CVPR. pp. 1–6 (2007)
3. Chaudry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for recognition of human actions. In: CVPR. pp. 1932–1939 (2009)
4. De Cock, K., Moore, B.D.: Subspace angles between linear stochastic models. In: CDC. pp. 1561–1566 (2000)
5. Doretto, G., Chiuso, A., Wu, Y.N., Soatto, S.: Dynamic textures. Int. J. Comput. Vision 51(2), 91–109 (2001)
6. Martin, R.J.: A metric for ARMA processes. IEEE Trans. Signal Process. 48(4), 1164–1170 (2000)
7. Sohail, A.S.M., Rahman, M.M., Bhattacharya, P., Krishnamurthy, S., Mudur, S.P.: Retrieval and classification of ultrasound images of ovarian cysts combining texture features and histogram moments. In: ISBI. pp. 288–291 (2010)
8. Spencer, J.K., Adler, R.S.: Utility of portable ultrasound in a community in Ghana. J. Ultrasound Med. 27(12), 1735–1743 (2008)