基于流形学习和混合模型的视频摘要方法

翟素兰 罗 斌 张春燕
(安徽大学计算智能与信号处理教育部重点实验室,合肥 230039)

摘 要 视频摘要是进行视频浏览、视频检索、视频索引等视频应用的前提,而且视频摘要类似于文本的摘要,也是对视频内容的一个简短概括。为了自动获得既包含视频的主要信息,而冗余信息又少的视频摘要,提出了一种基于流形学习和有限混合模型的自动视频摘要方法。该方法通过对视频序列进行流形建模,首先得到视频场景的初次分割;然后对包含内容较多的场景,使用等距降维方法计算视频帧的特征向量;最后将视频帧的特征向量输入到混合模型进行聚类分析,得到更细粒度的摘要结果。为了实现视频摘要的自动处理,所采用的混合模型需要具有模型选择功能。混合模型的聚类结果和流形建模的结果共同构成了视频摘要。视频分割片段的实验结果表明,在不需人为干预的情况下,所提供的视频摘要不仅包含视频主要内容,而且冗余信息少。

关键词 视频摘要 流形学习 等度降维 模型选择 混合模型
中图法分类号:TP391.3 文献标识码:A 文章编号:1006-8961(2008)04-0735-06

Video Abstraction Based on Manifold Learning and Mixture Model

ZHAI Su-lan, LUO Bin, ZHANG Chun-yan
(*Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University, Hefei* 230039)

Abstract Techniques for video abstraction has attracted tremendous attention for its application in video browsing, video indexing, video retrieval and so on. Video abstraction is brief summary of the video content like the text abstraction. In the paper, an automatic method for video abstraction is presented which is based on manifold modeling and mixture model. Manifold modeling is applied to generate the scene manifold of the video, Isomap is used to reduce the dimension of the video frames in larger scenes and the low dimension vectors are put into the mixture model with model selection to complete cluster analysis. Because mixture model with model selection can adapt to the data from any distribution, it is applied to generate the video abstraction automatically. The results from manifold modeling together with those from mixture model constitute the abstraction results. The experiments present the abstraction with less redundance, which demonstrates the effective and efficiency of the proposed method.

Keywords video abstraction, manifold learning, isomap, model selection, mixture model

1 引 言

随着视频信息的剧增,各种应用随之而生。视频摘要是视频浏览、视频索引、视频检索等应用的前提,已吸引了国内外众多研究者的关注。视频摘要类似于文本内容的摘要,也就是对视频内容的简短总结。更具体一点来说,视频摘要就是一串静止或运动的图像,它是用精简的方式代表视频的内容,并保留视频的要点[1]。

视频帧是视频的最小单位,镜头是摄像头的一次连续启停拍摄得到的一组图像,场景是内容相近的镜头组。视频摘要的实质就是对视频进行场景分析,以便得到视频场景的代表帧。视频摘要方法主要有基于关键帧的视频摘要和基于视频整体内容的视频摘要两类。其中基于关键帧的视频摘要,需要

基金项目:国家自然科学基金项目(60772122);安徽省教育厅自然科学研究重点项目(KJ2007A045)
收稿日期:2007-03-27;**改回日期**:2007-11-01

第一作者简介:翟素兰(1977~),女。讲师。现为安徽大学计算智能与信号处理博士研究生。主要研究方向为视频分析、模式识别。
E-mail:Sulan_zhai@tom.com

首先对视频进行镜头分割,然后抽取视频的关键帧,最后通过对视频关键帧进行聚类来得到视频摘要,但这种方法需要进行镜头的分割,而目前基于内容的镜头分割还存在一些问题;另一种就是将视频作为一个整体,先对整个视频通过一些聚类方式等来找出视频的场景,然后使用场景的代表帧序列作为视频的摘要[2,3]。无论是基于镜头的摘要,还是将视频作为一个整体的摘要,都需要进行特征提取和聚类方法的选择。基于聚类的方法,一般都需要首先进行低级特征的提取,如颜色、形状、纹理等;然后对特征矩阵进行降维;最后选取合适的聚类方法对数据进行聚类分析。经过后处理,即可得到视频摘要的结果。目前采用的降维方法,主要是线性降维方法如主成分分析(PCA)等,但线性方法无法发现非线性数据的特征维数。摘要的结果基本上是通过一步聚类得到,但是由于视频数据内容复杂,通过一步聚类很难产生满意的摘要结果,因此往往需要多种方法相互结合,才能产生粗细相辅的摘要结果。

2 基于流形建模和混合模型的视频摘要

视频摘要,其追求的目标是:摘要的内容要涵盖视频的主要信息,不仅要冗余信息少,而且要尽量实现自动化或半自动化,以减少人的干预[4]。视频摘要难点主要是视频数据的复杂性和先验知识的缺乏。本文提出了一种自动的视频摘要方法,即首先使用流形对视频序列进行建模,同时将视频的场景表示成不同的流形,以得到视频的粗粒度的摘要;之后使用等度降维(ISOMAP)进行非线性降维,并将降维的结果输入到具有模型选择功能的混合模型中进行聚类分析,而聚类的结果和部分流形的内容就共同构成视频的摘要(如图 1 所示)。

2.1 视频场景的流形建模

认知科学发现,人的面部表情可以用内嵌在高维空间的低维流形来表示[2,5]另外,手的姿势也具有低维流形表示。但是数据往往来自多个流形,显然人的面部表情和姿势就来自不同的流形。由于不同的流形之间存在间隙[6],因此如果使用能够保持流形本质关系的局域邻接图来表示来自不同流形的数据,则这个邻接图就不会是一个完全连通图,而是由若干连通分量组成。

由于不同场景视频帧的内容差异性大,所以是来自不同的流形。本文提出通过构造视频的邻接图来得到视频场景的流形表示。

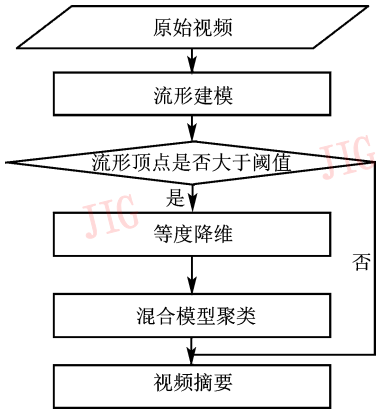

图 1 本文方法的流程图

Fig. 1 Diagram of the video abstraction method

给定视频 $\boldsymbol{V}$ 是由视频帧序列 $\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\}$ 构成的,其场景流形构成过程如下:

(1) 计算视频帧的距离 $d(i,j)=\|\boldsymbol{f}_i-\boldsymbol{f}_j\|$,$\boldsymbol{f}$ 是长度为 $M\times N$(M,N 是视频帧的行、列数)的 1 维向量,其元素是视频帧相应像素的灰度值;

(2) 构造 K 近邻图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$,$V$ 是视频的所有帧,$\boldsymbol{E}\subseteq\boldsymbol{V}\times\boldsymbol{V}$,$(\boldsymbol{f}_i,\boldsymbol{f}_j)\in\boldsymbol{E}$,如果 $\boldsymbol{f}_i$ 是 $\boldsymbol{f}_j$ 的 K 近邻或 $\boldsymbol{f}_j$ 是 $\boldsymbol{f}_i$ 的 K 近邻;

(3) 搜索图 $\boldsymbol{G}$ 的连通分量及与连通分量 $\boldsymbol{G}_i$ 对应视频场景 $\boldsymbol{S}_i$。

2.2 视频场景的特征降维

在使用混合模型分析包含视频帧较多的场景 $\boldsymbol{S}_i$ 之前,需要选取有效的降维方法来进行特征降维。目前主要的降维方法有主成分分析、多维尺度分析(MDS)等。虽然这些降维方法简单、易行,但只能发现线性高维空间或近似线性高维空间的真实结构。非线性降维方法是一个新的研究方向,其成果主要有局部线性嵌入(locally linear embedding,LLE)[7]和等度降维(Isometric mapping ISOMAP)[8]。ISOMAP 先使用近邻图的最短路径作为近似测地线距离,用于代替不能表示内在流形结构的欧氏距离;然后将其输入到多维尺度分析中进行处理,以便发现嵌入在高维空间的低维坐标。ISOMAP降维方法已经成功地用于人脸、手写体、姿态等图像数据的降维。

利用与场景 $\boldsymbol{S}_i$ 对应的 K 近邻图 $\boldsymbol{G}_i$ 来对场景 $\boldsymbol{S}_i$ 降维由以下两步构成:

(1) 用 Dijkstra 通过计算图 $\boldsymbol{G}_i$ 上两点 i,j 间的最短路径 $d_{\boldsymbol{G}_i}(i,j)$ 来作为场景流形 $\boldsymbol{S}_i$ 的近似测地线距离 $d_{\mathrm{M}}(i,j)$;

(2) 应用 MDS 算法来构建 d 维 Euclidean 空间 Y 上的低维嵌入[8]。

ISOMAP 的效果在仿真实验和自然视频数据等方面都得到了验证[9]。对视频场景 S_i 降维处理后就得到了场景的特征数据集:n_i 为场景 $\boldsymbol{S}_i$ 包含的视频帧数,d 维属性,记作 $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{n_i}\}$。

2.3　视频场景的混合模型分析

对于包含视频帧较多的视频场景 $\boldsymbol{S}_i$,为了得到更细粒度的视频摘要结果,需要进一步进行聚类分析。由于有限混合模型可以通过选择合适分量来对异常复杂的数据分布进行建模,又由于视频数据的复杂性和视频摘要自动化的需要,故本文采用了 Figueiredo 和 Jain 提出的一种基于类似 MDL(minmum description length)准则的混合模型方法[10],该方法可以将模型选择和参数估计无缝地接合起来,并显示了该方法优于大多数的准则。

场景 $\boldsymbol{S}_i$ 的特征数据集 $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{n_i}\}$ 的似然对数函数为

$$L(\boldsymbol{\Theta}_k, \boldsymbol{Y}) = \log \prod_{i=1}^{n_i} \sum_{m=1}^{k} a_m p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_m) \quad (1)$$

式中,a_m 是混合比,k 是高斯分量的个数,$\sum_{m=1}^{k} a_m = 1$,参数 $\boldsymbol{\Theta}_k \equiv \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k, a_1, \cdots, a_{k-1}\}$,$p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_m)$ 为高斯分量密度函数。

基于类似 MDL 的混合模型中,参数的估计采用最大似然估计 $\hat{\boldsymbol{\Theta}}_k = \operatorname{argmax}(L(\boldsymbol{\Theta}_k, \boldsymbol{Y}))$ 作为参数 $\boldsymbol{\theta}$ 的估计值。具体可通过 EM 算法实现。实现这个基于类似 MDL 的混合模型参数估计的 EM 算法分以下两步:

E 步

$$Q(\boldsymbol{\Theta}_k, \hat{\boldsymbol{\Theta}}_k^{(t)}) = \mathrm{E}\{L_{\mathrm{C}}(\boldsymbol{\Theta}_k, \boldsymbol{Y}, z) \mid \boldsymbol{Y}, \hat{\boldsymbol{\Theta}}_k^{(t)}\} \quad (2)$$

这一步也就是计算似然对数 L_{C} 的条件期望,由于 $z_m^{(i)}$ 是二值的,因此可得

$$\mathrm{E}\{z_m^{(i)} \mid \boldsymbol{Y}, \hat{\boldsymbol{\Theta}}_k^{(t)}\} = \frac{\hat{a}_m^{(t)} p(\boldsymbol{y}_i \mid \hat{\boldsymbol{\Theta}}_k^{(t)})}{\sum_{j=1}^{k} \hat{a}_j^{(t)} p(\boldsymbol{y}_i \mid \hat{\boldsymbol{\Theta}}_k^{(t)})} \equiv w_m^{(i,t)}$$

这与标准的 EM 算法的 E 步相同。

M 步,更新参数

$$\hat{\boldsymbol{\Theta}}_k^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}_k} \{Q(\boldsymbol{\Theta}_k, \hat{\boldsymbol{\Theta}}_k^{(t)}) + \log p(\boldsymbol{\Theta}_k)\} \quad (3)$$

$p(\boldsymbol{\Theta}_k, k) = p(\boldsymbol{\Theta}_k) p(k)$,这里 $p(\boldsymbol{\Theta}_k)$ 是 $p(\boldsymbol{\Theta}_k \mid k)$ 的缩写;t 是迭代次数。M 步参数的更新可将 MDL 准则无缝地融入到算法中。MDL 准则为

$$\hat{\boldsymbol{\Theta}}_k = \operatorname*{argmin}_{k, \boldsymbol{\theta}_k} \left\{ \frac{\log |I(\boldsymbol{\Theta}_k)|}{2} - L(\boldsymbol{\Theta}_k, \boldsymbol{Y}) - \log p(\boldsymbol{\Theta}_k) \right\} \quad (4)$$

其中,$I(\boldsymbol{\Theta}_k) = \mathrm{E}(-\nabla^2_{\boldsymbol{\theta}_k} L(\boldsymbol{\Theta}_k, \boldsymbol{Y}))$

$$a_m^{(t+1)} = \frac{\max\left\{0, \left(\sum_{i=1}^{N} w_{m,i}^{(t)}\right) - N/2\right\}}{\sum_{j=1}^{k} \max\left\{0, \left(\sum_{i=1}^{N} w_{j,i}^{(t)}\right) - N/2\right\}} \quad (5)$$

$\boldsymbol{\Theta}_k$ 的估计同标准的 EM 算法。式(5)在估计 a_m 的过程中,不但消除了一些不满足条件的分量,并且使计算不会陷入局部极小。实验证明,该方法可以估计出数据的真实模型数。

$k_{\max}$ 在本文中选取为 $\sqrt{n_i}$(n_i 是场景 $\boldsymbol{S}_i$ 中包含的视频帧数)。由于基于近似 MDL 准则的混合模型可以选择模型的阶,因此将场景 $\boldsymbol{S}_i$ 的特征向量输入到混合模型中,就可以自动得到场景更细的视频摘要结果。

3　实验分析

为了验证本文算法的效果,在 Matlab 环境下,从开放视频组织的视频库中[11]随机选择了 40 幅视频进行了实验,并将得到的视频摘要结果与开放的视频组织提供的故事板进行了比较。这个组织提供的视频故事板是使用 C-means 方法得到的,并且经过了必要的人的干预。本文实验主要包括以下 3 部分:(1) 在构造视频的流形表示中,K 值的选择问题;(2) 视频的流形表示;(3) 获取视频摘要。由于视频数据噪声较大,同时为了减小计算量,因此对于大小为 320 × 240 的视频帧,不是直接转化为 240 × 320 的 1 维向量,而是先将视频帧进行 4 × 4 的分块化处理,然后用块的灰度均值代替整个块的灰度,并将其转化为大小 80 × 60 的 1 维向量。

3.1　参数 K 的选择

视频场景的流形表示是通过构造 K 近邻图来实现的,因此 K 值的选择问题似乎是一个重要的问题。本文从 indi001 视频序列中选择了连续的 600 帧视频进行 K 近邻图的构造。由图 2 可见,当 K 趋于无穷大时,任何数据总可以构成一个完全连通的 K 近邻图。实际上,应该选取曲线的拐点处的 K 值。本实验取 $K = 5$。但是由于视频数据的差异性较大,因此不同的视频需要取不同的 K 值。

虽然较大的 K 值会使一些类似的场景合并，但由于视频摘要方法会对合并的场景通过聚类进行细分，所以本文实验取 $K=10$，这对实验结果没有影响。

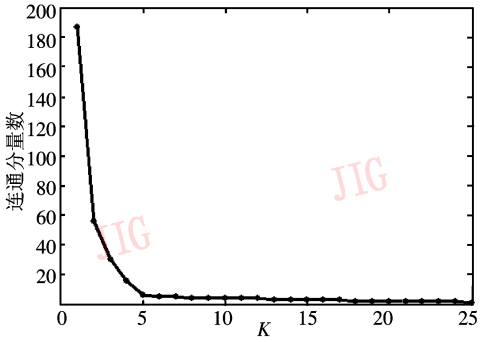

图 2 连通分量数与 K 的关系

Fig. 2 Number of components of the KNNG as a function of K for the video indi001

3.2 流形与视频场景

为了验证视频场景与流形的关系，本文选择 hall. avi 和 indi001. avi 两个视频进行实验。hall. avi 视频片段属于同一个场景，而 indi001. avi 视频的前 600 帧却包含了多了场景。根据本文提出的方法，将对这两个视频片段，首先构成一个完全连通的近邻图；然后使用 ISOMAP 降维。实验中，hall. avi，$K=3$，indi001. avi 前 600 帧视频片段，$K=18$。而且选取的目标维数要求保证信息的损失不超过 10%[10]。大部分视频目标维数为 4 就可以满足要求，但考虑到视频信息的差异性，本实验目标维数选为 6，并且选取前 3 维显示（如图 3、图 4 所示）。

可见，属于同一场景的视频帧来自一个流形或流形附近；而不同的场景，由于视频帧的区别大，因此来自于不同的流形。

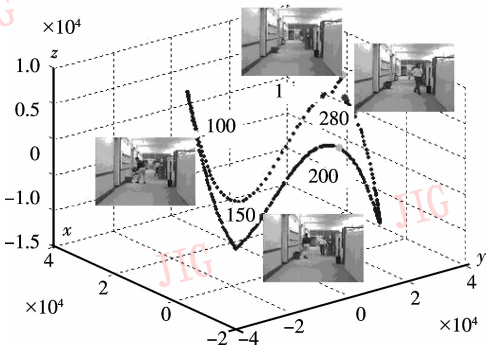

图 3 Hall 的流形表示

Fig. 3 Manifold representation of hall

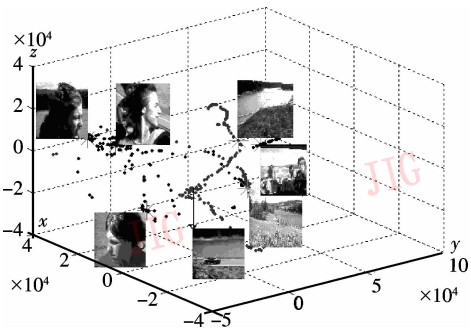

图 4 Indi001 片段的流形表示

Fig. 4 Manifold representation of indi001section

3.3 视频摘要结果

本文对 40 幅真实视频进行了实验。实验时，采用本文方法，首先构造视频的 K 近邻图，并对视频进行首次的场景分割；然后对包含视频帧较多的场景使用 ISOMAP 降维，同时将降维的结果输入到混合模型中进行聚类分析；最后将距离流形中心或混合模型聚类的类中心最近的视频帧作为代表帧，并按时间进行排序，作为视频摘要结果。实验中，仅对最大的场景，且包含视频帧多于 300 的视频进行聚类分析。

Indi001 视频包含场景较多，两个方法得到的摘要相比，本文方法得到的是 19 个代表帧（如图 5 所示），而开放的视频组织提供的故事板是 20 个代表帧（如图 6 所示）。其中 15 个基本相同。图 5 的第 4 行第 1 列代表帧，故事板（图 6）中没有，但这一帧有着较丰富的语义信息，而对于故事板（图 6）中的

图 5 本文算法得到的 indi001 视频摘要结果

Fig. 5 Indi001 abstraction by the proposed method

第 5 行第 1 列代表帧,本文方法得到的视频摘要结果(图 5)中没有,但可以在其他的代表帧中找到相关信息。可见,本文方法得到的视频摘要结果包含的语义内容较完整。

图 6　Open-video 提供的 indi001 视频摘要结果

Fig. 6　OV storyboard for indi001

bor04_011 的两个视频摘要结果相比,本文算法得到的视频摘要结果是 5 个聚类(如图 7 所示),比故事板(如图 8 所示)多了一个,但这一个不是冗余帧。

图 7　本文算法得到的 bor04_011 视频摘要

Fig. 7　Bor04_011 abstraction by the proposed method

图 8　Open-video 提供的 bor04_011 视频摘要

Fig. 8　OV storyboard for bor04_011

本文算法得到的 ugs02_002 的视频摘要结果为 7 个代表帧(如图 9 所示),而故事板则有 8 个代表帧(如图 10 所示)。故事板包含了 1 个明显的冗余帧,即图 10 中的第 4 和第 5 幅只需 1 幅。图 9 中第 2 幅,是将距该场景流形中心最近的帧作为代表帧,而故事板中则把这个场景分成了两类,即图 10 中第 2 和第 3 幅。这又是一个冗余的帧。对于海边塔楼这个场景,尽管两种方法得到的视频摘要结果有差异,但从视觉内容上,本文算法得到的视频摘要结果不是冗余帧。

图 9　本文算法得到的 ugs02_002 视频摘要

Fig. 9　Ugs02_002 abstraction by the proposed method

图 10　Open-vido 提供的 ugs02_002 视频摘要

Fig. 10　OV storyboard for ugs02_002

由图 11、图 12 可以看出,两种方法得到的 wth_02 视频摘要结果是相同的。

图 11　本文算法得到的 wth_02 视频摘要

Fig. 11　Wth_02 abstraction by the proposed method

图 12　Open-video 提供的 wth-02 视频摘要

Fig. 12　OV storyboard for wth_02

通过对两种方法得到的视频摘要结果进行比较总结可见,本文方法的视频摘要结果与开放视频组织提供的故事板的关系分为以下 4 类:(1) 两种结果代表帧相同;(2) 少于故事板的;(3) 多于故事板

的;(4) 代表帧不匹配的(如表 1 所示)。不难得出以下结论:本文提供的方法不仅可以完全自动地进行视频摘要,而且与经过一定干预的故事板相比,不仅包含了视频的主要信息,且冗余帧少。

表 1 实验结果总结

Tab. 1 Summary of experiment results

总视频片段数	各类视频片段数			
	相同	少于	多于	不匹配
40	3	28	5	4

4 结 论

本文提出了一种基于流形学习和混合模型结合的自动视频摘要方法。实验结果表明,本文方法获得的视频摘要结果不仅包含内容较完整,且冗余信息少。该方法使用流形表示一个场景,并采用流形中心表示这个流形,同时选取距流形中心最近的视频帧作为代表帧。而如何使用一个更加完整的形式表示这个流形则是一个需要继续研究的问题

参考文献(References)

1 Dimitrova N, Zhang H J, Shahray B, *et al*. A applicaitions of video content analysis and retrieval[J]. IEEE Transactions on Multimedia, 2002, **9**(3):42 ~ 55.

2 Ngo C W, Ma Y F, Zhang H J. Video summarization and scene detection by graph modeling[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2005, **15**(2):96 ~ 305.

3 Mundur P, Rao Y, Yesha Y. Keyframe-based video summarization using Delaunay Clustering[J]. In: International Journal on Digital Libraries, 2006, **6**(2):219 ~ 232.

4 Sebe N, Lew M S, Smeulders A W M. Video retrieval and summarization [J]. Computer Vision and Image Understanding, 2003, **92**(2): 141 ~ 146.

5 Weinberge K Q, Saul L K. Unsupervised learning of image manifolds by semidefinite programming[A]. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR2004)[C], Washton, USA, 2004, **2**:988 ~ 995.

6 Souvenir R, Pless R. Manifold clustering[A], In: Proceddings of International Conference on Computer Vision (ICCV2005)[C], Beijing, China, 2005, **1**:648 ~ 653.

7 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, **290**(5500):2323 ~ 2326.

8 Tenenbaum J B, Silva V de, Langford J C. A Global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, **290**(5500):2319 ~ 2323.

9 Pless R. Using isomap to explore video sequences[A]. In: Proceedings of International Conference on Computer Vision (ICCV2003)[C], Nice, France, 2003, 1:1433 ~ 1440.

10 Jain A K, Figueiredo M A F. Unsupervised selection and estimation of finite mixture model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, **24**(3):381 ~ 396.

11 Open_Video Projec[EB/OL]. http://www. open_video. org. 2007-03-20.