ISSN 1000-0054 清华大学学报 (自然科学版) 2005 年 第 45 卷 第 1 期 14/37
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2005, Vol. 45, No. 1 52-56

语音合成中基于听辨指导的权重训练算法

吴志勇, 蔡莲红, 蔡 锐

(清华大学 计算机科学与技术系, 普适计算教育部重点实验室, 北京 100084)

摘 要: 针对语音合成的基元选取中权重设定的问题提出了一种基于人工听辨指导的权重自动训练的方法。该方法首先通过人工听辨对现有的基元选取结果进行评测打分, 然后采取韵律逼近的方法对人工评测的结果进行学习, 进而对权重进行调整修正, 从而实现权重的自动训练。实验表明: 该方法较好地解决了权重设定的问题, 使得合成语音的自然度听辨得分由 3.49 提高到 4.02。同时, 该方法还使得语音合成系统在使用过程中根据用户反馈自动进行优化成为可能。

关键词: 语音合成; 文语转换; 基元选取; 权重训练

中图分类号: TN 912.33 **文献标识码**: A

文章编号: 1000-0054(2005)01-0052-05

Perceptual evaluation weight training for Text-to-Speech synthesis

WU Zhiyong, CAI Lianhong, CAI Rui

(Key Laboratory of Pervasive Computing of Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: A weight training approach was developed based on perceptual evaluations for unit selection in concatenative Text-to-Speech (TTS) synthesis. The speech units are selected based on current weights, with the synthesized results evaluated syllable by syllable by a perceptual listening test. The result gives new candidate units with perceptual scores. Then the weights are modified to match the new candidates to provide supervised weight training. Tests demonstrate that the method improves the naturalness of the synthesized result from about 3.49 to 4.02. The method also allows automatic optimization of the TTS system according to the user-behavior feedback.

Key words: speech synthesis; Text-to-Speech; unit selection; weight training

在基于大规模语音数据库的文语转换系统(Text-to-Speech, TTS)中, 通过韵律代价函数从语音数据库中优选语音基元。韵律代价函数被定义为若干韵律特征参数分代价的加权和。权重的设定对基元选取乃至最终的语音合成效果有很大影响。

文[1, 2]中利用权重空间的遍历和线性回归的方法对权重进行训练。Alias 等基于遗传算法对随机生成的权重样本进行筛选[3]。Park 等将基元选取看作模式识别中的分类问题, 采取语音识别中区分训练的方法训练权重[4]。上述方法在训练时忽略了作为主体的人的作用。事实上, 语音合成的目标是要尽量满足人的听觉要求, 因此, 有必要从听觉的知觉感受角度进行权重训练的研究。初敏等对拼接代价的设计从知觉角度进行了探讨, 研究了拼接代价与 MOS(mean opinion score)得分的关系[5], 对于权重的训练则未作详细的说明。

本文在权重训练时引入人工听辨的指导, 让系统根据人工听辨结果自动进行权重的训练。训练的目标是使得机器自动选音结果与人工听辨结果间的韵律距离达到最小, 从而使得语音基元的选取更趋于人的知觉过程。训练的过程基于当前语音库中的基元, 无需准备额外的训练语料, 而且更可在合成的过程中持续进行训练, 因而合成系统可以在使用的过程中根据用户反馈不断进行自动优化。

1 韵律代价函数

基元选取的目标是根据待合成文本的韵律上下文从语音数据库中选出韵律特征最为匹配的语音基元。本文利用韵律代价函数对此加以描述。

假设待合成语句中音节数为 n, 韵律代价函数定义为合成语句中所有音节的韵律匹配代价之和

$$C = \sum_{j=1}^{n} V(j), \tag{1}$$

其中 $V(j)$ 为第 j 个音节的韵律匹配代价, 反映了该音节候选语音基元的韵律特征参数与目标韵律特征

收稿日期: 2003-12-15
基金项目: 国家自然科学基金资助项目 (60275014)
作者简介: 吴志勇(1977-), 男(汉), 江苏, 博士研究生。
通讯联系人: 蔡莲红, 教授, E-mail: clh-dcs@tsinghua.edu.cn

参数间的匹配程度。假设待合成音节的目标韵律特征参数向量为 $X_j=(x_{j1},\cdots,x_{ji},\cdots,x_{jp})$，相应的候选语音基元的韵律特征参数向量为 $Y_j=(y_{j1},\cdots,y_{ji},\cdots,y_{jp})$，定义韵律匹配代价为

$$V(j)=\sum_{i=1}^{p}w_iV_i(j)=\sum_{i=1}^{p}w_iV_i(x_{ji},y_{ji}),\quad(2)$$

其中 $V_i(j)=V_i(\cdot,\cdot)$ 表示候选基元的第 i 个韵律特征参数和目标韵律特征参数间的匹配程度。对于韵律特征参数, 本文使用 32 维的参数向量加以描述, 包括: 音段参数 C (7 维)、位置参数 P (8 维)、韵律参数 S (5 维) 和关联参数 G (12 维)[6]。式 (2) 中 w_i 是反映第 i 个韵律特征参数对整体韵律匹配的影响因子的权重, 也即本文的考察对象。

2 基于听辨指导的权重训练

韵律代价函数中权重的设定既可借助专家知识及人为经验, 经过大量测试人工确定; 也可采用回归训练、神经网络等机器学习的方法自动进行训练。本文结合人为经验和机器学习的特点, 提出了一种基于人工听辨指导的机器学习的权重训练算法。

2.1 基于听辨指导的权重自动训练

基于听辨指导的权重自动训练, 采取人工听辨加机器自动学习的方法对权重进行调整。如图 1 所示, 权重训练的过程分为两大步:

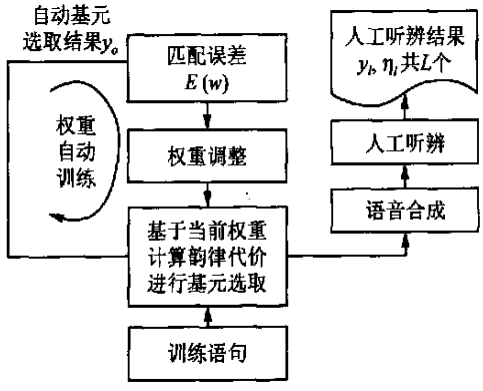

图 1 基于听辨指导的权重自动训练框架

1) 进行人工听辨实验, 给出听辨得分。首先由合成系统基于当前的权重对所有参加训练的语句进行基元选取及拼接合成, 然后由听辨人基于语句整体的合成效果对句中的每个音节进行听辨实验。若听辨人认为当前机器自动选音的结果 y_o 不合适, 则由听辨人给出其认为合适的候选基元, 即人工听辨结果 y_l, 同时给出听辨得分 η, 得分最高为 1, 最低为 0。听辨人可以给出多个听辨结果 $l=1,2,\cdots,L$。当只有一个听辨结果时, 自动给定默认得分 1。对于听辨人认为当前自动基元选取结果已经合适的音节, 算法自动默认该自动基元选取结果为人工听辨结果, 同时设置得分为默认值 1。

2) 进行机器学习和权重的自动训练。算法根据当前自动基元选取的结果以及若干人工听辨的结果进行学习和训练, 自动对权重进行调整和修正。训练的目标是使得自动基元选取结果与人工听辨结果间的匹配误差 $E(w)$ (式 (3)) 达到最小。

2.2 算法描述

假设参与训练的语句所包含的全部音节为 $\{x_1,x_2,\cdots,x_n,\cdots,x_N\}$, N 为全部音节的个数。对音节 x_n, 假设语音数据库中有 M_n 个候选的语音基元 $\{y_{nm}\,|\,m=1,2,\cdots,M_n\}$, 其中有 L_n 个人工听辨的结果 $\{y_{nl},\eta_l\,|\,l=1,2,\cdots,L_n\}$, 以及基于当前权重的自动基元选取结果 $\{y_{no}\}$。每个基元 y_{nm} 以 p 维的韵律特征参数向量加以表示。

定义匹配误差

$$E(w)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{L_n}\sum_{l=1}^{L_n}\eta_lV(y_{nl},y_{no}).\quad(3)$$

权重训练的目标是使得该匹配误差达到最小, 其反映了自动基元选取结果与人工听辨结果之间的韵律代价距离, 其中 $V(y_{nl},y_{no})$ 表示 y_{nl} 与 y_{no} 之间的韵律匹配代价, 与公式 (2) 类似, 其定义为各韵律特征参数匹配程度 $V_i(\cdot,\cdot)$ 的加权和

$$V(y_{nl},y_{no})=\sum_{i=1}^{p}w_iV_i(y_{nl},y_{no}).\quad(4)$$

假设初始权重向量 $w^0=(w_1^0,\cdots,w_i^0,\cdots,w_p^0)$ 经过 k 次训练后为 $w^k=(w_1^k,\cdots,w_i^k,\cdots,w_p^k)$, p 为权重向量的维数, 也即韵律代价函数中所考虑的韵律特征参数的个数。由于权重等比例缩放不影响基元选取的结果, 同时为了使权重的调整在整个向量空间上保持均衡, 并保证权重训练的收敛性, 引入归一化约束条件

$$\sum_{i=1}^{p}w_i^k=1,\quad w_i^k\geq 0.\quad(5)$$

进行第 $k+1$ 次训练时, 通过如下公式修正权重

$$w_i^{k+1}=\sigma(w_i^k+\delta^k\cdot q_i^k),\quad i=1,2,\cdots,p.\quad(6)$$

其中: σ 是为满足归一化的约束条件而引入的归一化因子, δ^k 为权重空间搜索的步长, q_i^k 为权重修正的具体方向及变化值。

权重训练的目标是使得式 (3) 定义的匹配误差达到最小, 该匹配误差为权重向量 w 的函数。由梯度下降法, 权重的调整应沿着 $E(w)$ 所定义的曲面

变化最陡的反方向进行。对于第 $k+1$ 次训练,权重修正的 q_i^k 为

$$q_i^k = -\frac{\partial E(\mathrm{w}^k)}{\partial w_i}. \tag{7}$$

由于匹配误差 $E(\mathrm{w})$ 并非权重向量 w 的显式函数,因此直接使用上式对 q_i^k 进行计算比较困难。本文通过如下方法计算 q_i^k。

2.3 权重的调整

权重的调整需要考虑到训练数据中韵律特征参数的分布特性。为此,统计训练数据中各个候选语音单元 y_{nm} 与人工听辨结果 y_{nl} 的韵律匹配代价 $V_i(y_{nm}, y_{nl})$ 相对于当前自动基元选取结果 y_{no} 的韵律匹配代价 $V_i(y_{nm}, y_{no})$ 的变化情况:

$$q_{i,n}^k = \frac{1}{M_n}\sum_{m=1}^{M_n}\frac{1}{L_n}\sum_{l=1}^{L_n}\eta[V_i(y_{nm}, y_{nl}) - V_i(y_{nm}, y_{nm})] \tag{8}$$

式(8)为训练数据中第 n 个训练样本(音节)的所有候选语音基元的第 i 个韵律特征参数相对于人工听辨结果及基于当前权重的自动基元选取结果的韵律特征的变化。对式(8)进行分析:如果 $V_i(y_{nm}, y_{nl}) > V_i(y_{nm}, y_{no})$,说明从当前训练样本的候选语音基元的统计特性上来看,第 i 个韵律特征相对于人工听辨的结果具有较强的区分力,属于较关键的韵律特征,对权重的修正是正向的;如果 $V_i(y_{nm}, y_{nl}) < V_i(y_{nm}, y_{no})$,对权重的修正是负向的;而若两者相等,其对权重修正的贡献度为 0。

对训练数据中的所有样本(音节)进行统计,则 q_i^k 的计算公式为

$$q_i^k = \frac{1}{N}\sum_{n=1}^{N} q_{i,n}^k. \tag{9}$$

式(6)中为了满足权重的归一化约束条件而引入了归一化因子 σ,有

$$\sum_{i=1}^{p} w_i^{k+1} = \sum_{i=1}^{p}\sigma(w_i^k + \delta^k \cdot q_i^k) = 1, \tag{10}$$

$$\sigma = \frac{1}{1 + \delta^k\sum_{i=1}^{p} q_i^k}, \tag{11}$$

因此,权重修正的最终计算公式为

$$w_i^{k+1} = \frac{w_i^k + \delta^k \cdot q_i^k}{1 + \delta^k\sum_{i=1}^{p} q_i^k}. \tag{12}$$

其中:q_i^k 根据式(8)和式(9)计算得到;δ^k 为权重空间搜索的变化步长,用于控制权重训练时的收敛速度,一般取为 0~1 之间的常数。

2.4 算法分析

考察上述训练过程及算法描述可以发现,无论训练数据的规模多大,算法均可适应该数据规模进行相应的训练。这是由于式(8)考虑了语音库中当前音节候选语音基元的韵律统计特性,从而保证了即使在训练数据规模较小的情况下,训练结果仍可反映当前语音库中韵律特征的分布特性。这一特性为算法实现时的自由扩展提供了可能。例如本文在算法实现时,针对不同的数据规模分别进行在线训练与离线训练,以使得当前训练样本的局部特性与所有训练样本的总体效果达到平衡。

上述算法描述中,必须对每句参与训练的语句进行人工听辨实验,这势必会极度费时费工。实际上,人工听辨实验只有在当前训练的语句未被语音库的语料覆盖时才是必须的。也即语音库的原始语料可以用来进行权重训练。事实上,实验表明,这种措施对权重的训练是极为有效的,该方法与上述一般意义上的算法描述的区别仅仅在于:将原始录音语料在语音数据库中的候选基元信息加以记录,并作为惟一的人工听辨结果加以训练即可。而且,基于原始录音语料的权重训练,其结果是当合成时遇到原始录音语料的部分或全部时总是尽量选择相应的连续语音片断,从而可以提高语音合成的自然度。

3 实验及分析

在实验中,共收集了 7 000 多句自然语音数据,由说标准普通话的女性播音员录制,语料包含近 85 000 个音节,覆盖了汉语 417 种有调音节及多种语境和韵律特征的搭配关系。基于上述数据抽取了一个中大规模的语音数据库作为实验基础。

3.1 权重训练实验

权重训练的语句,来源于两大部分:其一为上述语音数据库中对应的 7 000 多句的原始录音语料;其二为在合成系统使用和测试的过程中经过人工听辨的语料,共 953 句。对原始录音语料,将其在语音数据库中对应的候选语音基元的信息加以记录,并作为惟一的人工听辨结果,相应的人工听辨得分设置为默认值 1。训练时,将两部分数据放在一起进行统一训练。

由于韵律特征参数[6]中当前音节的声母、韵母以及声调类型是保证正确发音的基础,选音的结果必定要满足该三个参数,因此不参与权重训练,对于其余 29 个参数,利用上述算法进行训练。训练曲线如图 2 所示,其中图 2a 给出了部分韵律特征参数权重在训练时的变化情况,而图 2b 为匹配误差 $E(\mathrm{w})$

的变化曲线。可以看出，训练时匹配误差呈持续下降的趋势，训练后匹配误差值为0.032，对全部数据共训练了402次。训练后各个权重的具体值及分布在图3中作了进一步的说明。

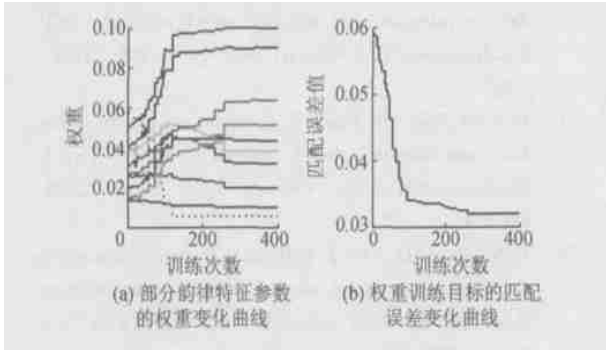

图2　权重训练变化曲线

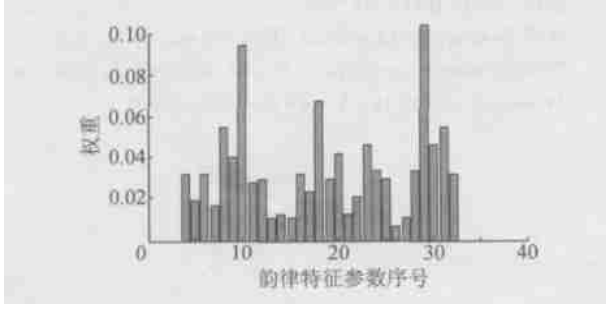

图3　权重训练结果

从图3可以看出，当前音节的基频特征(图中横坐标序号29)、位置信息(横坐标10/8/9)、韵律关联信息中的前音节基频特征、幅度(横坐标18/20)、后音节基频特征(横坐标23)、以及当前音节的重音属性(横坐标31)等参数权重较大，说明这些参数在基元选取中具有重要的作用。

3.2　音高曲线对比实验

为进一步验证权重训练对语音基元选取结果的影响，进行了选音对比实验。由于基频特征是影响语音自然度的重要因素，而且合成语音的韵律可以从音高曲线的变化反映出来，因此本实验主要对权重训练前后基元选取结果的音高曲线的变化进行分析比较。实验中选择50句待合成文本进行实验，经过对比分析，其中有41句的实验结果与自然语句的音高曲线更为接近，总比例为82%。

3.3　主观听辨实验

为从总体上进一步说明权重训练前后语音合成结果的自然度变化情况，设计进行了主观听辨实验。共100组(300句)语音用于听辨实验，每组语音包括权重训练前后各1句合成语音，以及进行比照的相应自然录音语音1句。实验时，将每组语音以随机的先后次序播放，要求听辨人根据自己的知觉感受给出其认为的自然度。自然度以5分制给出：5自然，4较自然，3可以接受，2较差，1不可接受。共10名听辨人参加了实验。

主观听辨实验的MOS得分结果如图4所示，图中给出了每名听辨人的MOS得分情况。经过权重训练后，合成语音自然度的MOS得分有较大提高：权重训练前(使用初始权重)平均MOS得分为3.49，经过本文的方法进行权重训练后平均MOS得分为4.02，比训练前平均提高了0.53。

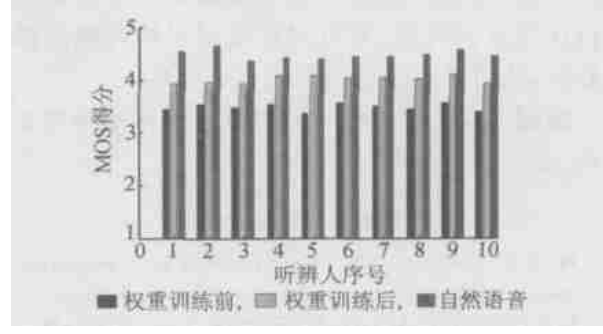

图4　主观听辨实验结果比较

3.4　分析与讨论

本文提出的基于人工听辨指导的权重自动训练算法具有如下一些特点：

1) 算法结合了人为经验和机器自动学习的优点：一方面，克服了传统的机器学习难于对学习过程加以人为指导的不足，通过人工听辨指导，使得权重自动训练的过程更为直观有效；另一方面，借助机器学习可以自动获取知识的优点，发现隐含在数据中的内在特性和规律，避免了完全基于人为经验的繁杂和片面性。

2) 权重训练的过程中，引入人工听辨的指导，从人的知觉感受出发直接对机器自动选音的结果加以评估，其目标是使得机器自动选音的结果与人工听辨的结果之间的韵律代价距离达到最小，从而使得语音基元的选取更趋于人的知觉过程。

3) 权重训练时，直接基于当前语音数据库中的语音基元对当前语料的合成结果进行评价，无需录制额外的训练语音，训练的过程更为便捷。

4) 算法的实现机制使得权重的训练可以在语音合成系统使用的过程中持续进行，使得合成系统在使用过程中根据用户的反馈自动进行优化成为可能。这种优化基于对用户行为习惯反馈的分析，可以满足特定用户的知觉习惯的要求，在使用过程中逐渐提高系统的性能和特定用户的知觉满意度。

4 结 论

基于韵律代价函数的基元选取中,各个韵律特征参数的权重对基元选取乃至语音合成的结果有较大影响。本文对权重的训练算法进行了研究,提出了基于人工听辨指导的权重自动训练算法,让系统根据人工听辨的结果自动进行权重的调整。实验表明,本文提出的算法结合了人为知觉感受和机器自动学习的优点,训练的结果使得合成语音的自然度有一定的提高。另外,本文提出的算法中人工听辨指导的引入,使得合成系统在应用中根据用户的反馈自动进行优化成为可能,并可对特定用户的习惯反馈进行分析,从而产生特定用户的自适应性。

致谢 本文中部分工作得益于和陶建华博士的讨论,在此表示感谢。

参考文献 (References)

[1] Hunt A J, Black A W. Unit selection in a concatenative speech synthesis system using a large speech database [A]. ICASSP96 [C]. Atlanta: IEEE Press, 1996. 373-376.

[2] Meron Y, Hirose K. Efficient weight training for selection based synthesis [A]. EuroSpeech99 [C]. Budapest: ISCA Press, 1999. 2319-2322.

[3] Alias F, Llora X. Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis [A]. EuroSpeech2003 [C]. Geneva: ISCA Press, 2003. 1333-1336.

[4] Park S S, Kim C K, Kim N S. Discriminative weight training for unit-selection based speech synthesis [A]. EuroSpeech2003 [C]. Geneva: ISCA Press, 2003. 281-284.

[5] PENG Hu, ZHAO Yong, CHU Min. Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation [A]. ICSLP2002 [C]. Denver: IEEE Press, 2002. 2613-2616.

[6] 吴志勇, 蔡莲红. 语音合成中的韵律关联模型[J]. 中文信息学报, 2004, **18**(2): 44-50.
WU Zhiyong, CAI Lianhong. Prosodic correlation model in Text-to-Speech synthesis [J]. *J Chinese Information Processing*, 2004, **18**(2): 44-50. (in Chinese)

(上接第 43 页)

5 结 论

本文给出了一种对带有抖动的视频序列进行稳定而得到平滑的视觉效果的方法,采用B 样条设计数字滤波器来进行运动滤波。前人的方法总是在能达到的平滑度和对拍摄的主观运动的跟踪上取折中,因此对于不同幅度抖动的视频产生的运动平滑效果不同,此外从频率角度出发的滤波器无法保证运动曲线的光滑度。本文中的滤波器规定了摄像机的主观运动方式(包括采用二次B 样条来模拟和相关时间长度设置)的同时也就限定了运动的平滑度,不需要设置抖动的参数,相关时间是跟人眼对运动观察的舒适程度相关的,比较稳定,而抖动的幅度和模式是多变的。理论分析和实验结果证明,在全局运动能够比较准确求得的条件下,基于B 样条滤波器对抖动具有很好的平滑效果。

本文的方法是应用在家用数码相机和摄像机上,不适用于表现剧烈的主观运动的场合,比如特技摄影。现在只对平动和平面转动进行了讨论,今后需要进一步扩展考虑摄像机变焦、绕光轴旋转等其他维度的运动。此外,对于全局运动的计算方法的准确度和复杂度还有很大的改进余地。

参考文献 (References)

[1] Uomori K, Morimura A, Ishii H, et al. Automatic image stabilizing system by full-digital signal processing [J]. *IEEE Transactions on Consumer Electronics*, 1990, **36**(3): 510-519.

[2] Egusa Y, Akahori H, Morimura A, et al. An electronic video camera image stabilizer operated on fuzzy theory [A]. IEEE International Conference on Fuzzy Systems [C]. San Diego: IEEE Press, 1992, 851-858.

[3] Ko S J, Lee S H, Lee K H. Digital image stabilizing algorithms based on bit-plane matching [J]. *IEEE Transactions on Consumer Electronics*, 1998, **44**(3): 617-622.

[4] Jin J S, ZHU Zhigang, XU Guangyou. A stable vision system for moving vehicles [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2000, **1**(1): 32-39.

[5] Litvin A, Konrad J, Karl W. Probabilistic video stabilization using Kalman filtering and mosaicking [A]. SPIE Conference on Electronic Imaging [C]. Santa Clara, 2003. 663-674.

[6] Eric W W. B-Spline—from Mathworld [EB/OL]. http://mathworld.wolfram.com/B-Spline.html.

[7] Tudor P N. MPEG-2 video compression [J]. *IEE Electronics & Communications Engineering Journal*, 1995, **7**(6): 257-264.

[8] Dufaux F, Konrad J. Efficient, robust, and fast global motion estimation for video coding [J]. *IEEE Transactions on Image Processing*, 2000, **9**(3): 497-501.